STATISTICAL METHODS IN HIGH-DIMENSIONAL STRUCTURED DATA

by

Lei Huang

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

March, 2016

Abstract

Modern data pose several challenges to statistical analysis. They are not only big in size, high in dimensionality but also complex in structure. For example, diffusion tensor imaging (DTI) measures water diffusion within white matter. A 3D DTI of the brain consists of millions of voxels with a complex structure induced by the anatomical shape of the human brain. The first project is dedicated to studying the association between DTI images, which are a proxy of effective connectivity in the brain, and clinical outcomes. To account for the spatial structure and to reduce the dimensionality, we propose a hierarchical Bayesian "scalar-on-image" regression method designed specially for aribitary-shaped image data. The second project is concerned with defining and implementing the methodology for fitting populations of separable spatio-temporal processes. Methods were motivated by and applied to functional magnetic imaging (fMRI) studies where large spatial images of the brain are observed at a dense grid of points over time. We show that separability combined with principal component analysis of latent processes provide a fast and feasible tool for dimensionality reduction and model-driven discovery. The third project is dedicated to

ABSTRACT

modeling matrix-valued data oberved repeatedly over time. The project was motivated by an accelerometry study which collected minute-level activity intensity data 24 hours a day and 7 days a week. Considering the week as a unit of measurement, the basic measurement unit is a 1440 (minutes within a day) by 7 (days within a week) dimensional matrix. As data are observed for many weeks for the same subject this induces a natural within-subject clustering structure. We use a linear mixed effect model to account for the multilevel design, while the 2D structure is handled via normal matrix-variate distribution. All three proposed methods are scalable and software is available.

Advisor:

Ciprian Crainiceanu, PhD

Committee:

Adam Spira, PhD (chair, SPH mental health)

Ciprian Crainiceanu, PhD (advisor, SPH biostatistics)

Alden Gross, PhD (SPH epidemiology)

Martin Lindquist, PhD (SPH biostatistics)

Alternates:

Jim Pekar, PhD (SOM radiology and radiological science)

Vadim Zipunnikov, PhD (SPH biostatistics)

Acknowledgments

First I would like to thank my advisor, Ciprian Crainiceanu. For my academic development, he has always been a constant source of advice. Whenever I have questions, his door is always open. He discussed with me on a detailed technical problem on PCA decomposition. He spent hours to go through word by word with me during the revision on my first paper. More importantly, he teaches me how to think strategically and independently towards a new problem. For my career development, he is very supportive. He encourages me to explore different possibilities and helps me to find what I really love to do. I know when I fail, he can be the last resort to seek advice. To me, Ciprian is not only a great advisor or mentor but also a great friend. I have learned from him how to be a good biostatistical researcher, and more importantly, how to be a great person.

Thanks to my thesis committee for their advice over the years: Jim Pekar, Adam Spira, Martin Lindquist, Alden Gross and Vadim Zipunnikov.

Thanks to Philip Reiss for bringing me into the biostatistics world and your great advice on my different projects.

ACKNOWLEDGMENTS

Thanks to all the SMARTies. It is so great to be part of the group. The group is like a big family and is always full of great ideas and projects.

Warmest thanks to the students of the biostatistics department who overlapped with my time here. Special thanks to Chen Yue, Shaojie Chen, Huitong Qiu, Detian Deng and Yuting Xu: you guys are so amazing that I could not imagine what my PhD life will be without you. I am grateful to the students that began the program before me for their help on every aspect of my research and career (especially Haochang Shou and Jeff Goldsmith).

Finally, I would say thanks to my family. Thanks to my parents who have been supporting me all the time on every decision I made. They do not speak English and might never know what I am writing here, but I am sure they can feel it on the other side of the planet.

Contents

Abstract **ii**

Acknowledgments **iv**

List of Tables **x**

List of Figures **xi**

1 Introduction **1**

1.1 Statistical challenges 2

1.2 Organizational overview 3

2 Bayesian Scalar-on-Image Regression **6**

2.1 Introduction 8

2.2 Methods 11

2.2.1 An ill-posed multiple linear regression 12

2.2.2 Fitting through penalization 13

2.2.3 Our proposal: Imposing an Ising prior on a latent 0/1 map 15

2.2.4 Advantages of the Ising prior 18

2.2.5 Estimation 19

2.2.6 Tuning parameters 20

2.2.7 DTI study 21

2.3 Results 23

2.4 Discussion 30

3 Two-way PCA for matrix-variate data **33**

3.1 Introduction 35

3.2 Separable models for two-way matrix data 40

3.2.1 Separable additive model 41

3.2.2 Separable multiplicative model 42

3.2.3 Separable hybrid model 44

3.3 Separable two-way matrix-variate PCA 45

3.3.1 Eigenvectors and eigenvalues 46

3.3.2 Model estimation 48

3.4 Two-way matrix data with white noise 52

3.5 Simulation 54

3.6 Data application 56

3.6.1 Variation analysis and principal component analysis 57

3.6.2 Distribution of PC scores 60

3.6.3 Association between component scores and pain rating 61

3.7 Discussion 63

4 Multilevel matrix-variate analysis **69**

4.1 Introduction 71

4.2 Model and Estimation 75

4.2.1 Covariance decomposition 77

4.2.2 Extension to more complex study designs 78

4.2.3 Principal component estimation 79

4.2.4 Principal score estimation 80

4.2.5 Data with white noise 81

4.3 Simulation 82

4.4 Data application 85

4.4.1 Eigenvalues and eigenfunctions 86

4.4.2 Score matrices and covariates 89

4.5 Discussion 93

5 Discussion and Future Work **96**

Appendices **100**

A1 Appendix to Chapter 4 100

Bibliography **108**

Curriculum Vitae **121**

List of Tables

2.1 MS patient characteristics. Disability data were obtained within 30 days of the MRI scan. 23

2.2 Prediction performance of the tuning parameter combinations for Figures 2.4–2.6 . The last two columns refer to mean square prediction error and proportion of variance explained. 27

3.1 Three types of separable spatio-temporal models. 45

3.2 Additional assumption for models with white noise to ensure identifiability of models (3.1), (3.2) and (3.3), respectively. 53

3.3 Average MSE of the principal components under different signal-to-noise ratio for hybrid model. 56

3.4 Estimated eigenvalues on temporal, spatial and multiplicative term. "percent var" stands for the percentage of variance explained by the component, and "cum percent var" means the cumulative percentage of variance explained 58

3.5 Coefficients for fixed effects for three mixed effect models which regress PCA scores on the pain rating scores. 63

4.1 Average MSE between estimated and true eigenvectors using 100 simulation data. 84

4.2 Average MSE between estimated and true scores using 100 simulation data. 84

4.3 Eigenvalues for PCA for accelerometry study 87

4.4 Models for association between events and between-subject scores. For the variable sex, female is the reference group and an asterisk indicates significance at level 0.05 92

List of Figures

2.1 Red region contains the rectangular region we use as a predictor of cognitive function. Background 3D brain is rendered from a T1 template image. . 10

2.2 Illustration of the multiple linear regression model, with cognitive disability measure as the scalar response and fractional anisotropy maps as the image predictor. 12

2.3 Histogram of the estimated coefficients with tuning parameter values $a = -3, b = 6, \sigma_\varepsilon^2 = 1.22, \sigma_\beta^2 = 0.05$, which were chosen by five-fold cross-validation. The middle bar refers to coefficients whose magnitude lies below a threshold of 0.01. The blue bars denote the coefficients which are less than -0.01 while the red ones denote the coefficients that are larger than 0.01. 24

2.4 The estimated coefficient images from Slice 7 to Slice 22. The estimation is overlaid on one single subject's FA scan image for anatomical reference. The tuning parameters are selected via cross validation, $a = -3, b = 6, \sigma_\varepsilon^2 = 1.22, \sigma_\beta^2 = 0.05$. Positive coefficients are shown in red, while blue denotes negative coefficients. The estimated mean square prediction error is 146.43 and the proportion of variance explained for predicted data is 22.00%. 25

2.5 The estimated coefficient images from Slice 7 to Slice 22 using tuning parameters $a = -2, b = 0.5, \sigma_\varepsilon^2 = 1, \sigma_\beta^2 = 0.03$. The estimation is overlaid on one single subject's FA scan image for anatomical reference. The estimated mean square prediction error is 150.25 and the proportion of variance explained for predicted data is 20.34%. 26

2.6 The estimated coefficient images from Slice 7 to Slice 22 using tuning parameters $a = -1, b = 0.5, \sigma_\varepsilon^2 = 0.775, \sigma_\beta^2 = 0.05$. The estimation is overlaid on one single subject's FA scan image for anatomical reference. The estimated mean square prediction error is 164.6 and the proportion of variance explained for predicted data is 19.22%. 27

2.7 Profile cross validation plot: in each panels, three of the tuning parameters are fixed at the values chosen by cross-validation while the remaining tuning parameter varies in the x-axis. The y-axis is the proportion of variance explained in the left-out data. 28

2.8 p-value map for voxel-wise linear regression fitting from Slice 7 to Slice 22. 30

3.1 The stimulation design for a single trial during a 46-second time course. . . 37

3.2 fMRI intensity measurements over 21 locations $\times$ 23 time points over two subjects. 37

3.3 The estimated principal components when N=500 in 100 simulations for hybrid model are shown in gray bands. Black curves are the true eigenvectors. The figure also contains the density function of the estimated σ_C^2, plotted with the red vertical line marking the true value. 66

3.4 Estimated overall, temporal and spatial mean functions for thermal pain fMRI data. The left panel shows the overall mean of fMRI across all subjects and all trials. The x-axis indexes the time course and the y-axis indexes the brain regions (see Table S.1 for more information). The middle panel shows the marginal mean of the temporal signal. The x-axis denotes the time course. The right panel shows the marginal mean of the spatial signal. 67

3.5 Principal components under hybrid model. The left panel shows the row-specific or spatial PCs. The right panel shows the column-specific or temporal PCs. The black, red, green and blue lines stand for the first, second, third and fourth components, respectively. For more information about locations see Table S.1. 67

3.6 Estimated principal component scores. Upper left panel: scatterplot of the 1st versus 2nd PC scores for the column-specific term. Other panels: distribution of the first few PC scores for column-specific term, row-specific term and multiplicative term versus stimulation setting covariate. 68

4.1 Log activity count plots for subject 1 and subject 2 across 10 weeks. 72

4.2 The figure illustrates the decomposition for a randomly chosen participant based on Equation (4.2). The left panel shows the deviation from the population mean $\mathbf{Y}_{ij} - \boldsymbol{M}$, the middle panel shows the estimated subject-specific deviation, $\mathbf{X}_i$, from the mean and the right panel shows the visit-specific deviation estimator, $\mathbf{W}_{ij}$, from the subject-specific mean. 76

4.3 Population mean of log activity counts within days and across days. 86

4.4 Principal components for $\mathbf{C}_X$, $\mathbf{R}_X$, $\mathbf{C}_W$ and $\mathbf{R}_W$, adjusted by eigenvalues. Black, red, green, blue lines are the 1st, 2nd, 3rd, and 4th principal components, respectively. 89

4.5 Boxplots of elements in between-subject score matrices grouped by event types. 90

4.6 Principal matrices which are the outer products between different principal components for the time and day dimension. 91

4.7 Violin plots for selected scores in within-subject score matrices in the first week, during event and after event. 94

A.1.1 Simulation result for signal-to-noise ratio $+\infty$. 104

A.1.2 Simulation result for signal-to-noise ratio 10. 105

A.1.3 Simulation result for signal-to-noise ratio 1. 106

A.1.4 Simulation result for signal-to-noise ratio 0.1. 107

Chapter 1

Introduction

1.1 Statistical challenges

Modern technology enables us to collect data much faster with ever increasing resolution. This increase in speed, size and complexity poses new challenges for statistical analysis. In this work, three main challenges are addressed.

The first challenge is dealing with various aspects of high dimensionality and small to moderate sample sizes. For example, brain MRIs contain millions of voxels while an imaging study may involve tens or hundreds of subjects. A common problem in such modern data sets is the so called "curse of dimensionality", which affects standard statistical techniques. For example, when regressing a subject-specific outcome (scalar) on an image (high-dimensional predictor), linear regression based on least squares minimization has an infinite number of solutions. Such problems can be addressed by constraining the solution space using additional assumptions. Two popular approaches for constraining the solution space are to impose L-1 or L-2 penalty on the regression coefficients, which is equivalent to assuming that coefficients are exchangeable and have a double-exponential or normal distribution, respectively. We focus on identifying solution restrictions or, equivalently, prior distributions on the model parameters, that are best suited for specific applications.

The second challenge is the complex data structure. For example, the shape of the brain is an irregular manifold. Thus, it becomes challenging to take into account spatial correlations when assessing the association between brain images and health outcomes. Another example of complex data structure that has become common are data in the form of two-

way matrices. Indeed, in fMRI one dimension corresponds to voxels or ROIs (space) while the other corresponds to time. In accelerometry studies, devices collect activity information 24 hours a day and 7 days a week for many weeks at a time. It is often convenient to organize data by weeks, as there is increasing evidence that activity during various days of the week is not exchangeable. In particular, there are major changes between week-days and week-ends.

The third challenge is the size of data. Indeed, motivated by the advance of the technology and reduced cost of storage, data sets are increasing dramatically in size.. This requires techniques that can easily scale up.

This research aims to develop new methodological tools designed to address these challenges.

1.2 Organizational overview

The first method is designed to quantify the association between a patient's disability score and their brain diffusion tensor image (DTI). DTI is a type of Magnetic Resonance Imaging designed to capture the degree of isotropy of water molecules. These images are expected to provide good proxy information about the effective, or anatomic, connectivity in the brain. Disruptions in anatomical connectivity are thus expected to be associated with adverse cognitive health outcomes. To study these associations, we propose a hierarchical Bayesian "scalar-on-image" regression approach. The approach uses an Ising prior to

restrict the solution space by allowing a latent binary map to indicate voxels that are predictive and maintains the spatial contiguity of predictive regions. The method is applied to a large study of association between fractional anisotropy estimated from DTI-MRI data at 198,250 voxels and cognitive disability in a cross-sectional sample of 135 multiple sclerosis patients.

The second method was motivated by multiple neuroimaging studies that acquire large spatial images of the brain observed sequentially over time. Such data are often stored in the form of matrices. To model these matrix-variate data we introduce a class of separable processes using explicit latent process modeling. To account for the size and two-way structure of the data, we extend principal component analysis to achieve dimensionality reduction at the individual level. We introduce necessary identifiability conditions for each model and develop scalable estimation procedures. The method is motivated by and applied to a functional magnetic resonance imaging study designed to analyze the relationship between pain and brain activity.

The third method is designed for modeling populations of repeated matrix-variate measurements. We use a linear mixed effect model to account for the multilevel design, while the 2D structure is handled via normal matrix-variate distributions. To achieve dimensionality reduction, we estimate and decompose the row- and column-specific covariance operators. The computational feasibility and performance of the approach is shown in extensive simulation studies. The method is motivated by and applied to a study that monitored physical activity of individuals diagnosed with congestive heart failure (CHF) over

a 3- to 10-month period. Two primary goals of the study were: 1) to quantify and model the long-term patterns of physical activity in individuals with CHF; and 2) evaluate the possibility of predicting adverse health effects via continuous activity monitoring.

Chapter 2

Bayesian Scalar-on-Image Regression

Abstract

Diffusion tensor imaging (DTI) measures water diffusion within white matter, allowing for in vivo quantification of brain pathways. These pathways often subserve specific functions, and damage to them may result in characteristic forms of disability. As a means of predicting clinical disability from DTI images, we propose a hierarchical Bayesian "scalar-on-image" regression procedure. Our procedure introduces a latent binary map that estimates the locations of predictive voxels, thereby resolving the ill-posed nature of the problem. By inducing a spatial prior structure, the procedure yields a sparse association map which also maintains spatial continuity of predictive regions. The method is demonstrated on a large study of association between fractional anisotropy and cognitive disability in a cross-sectional sample of 135 multiple sclerosis patients.

Keywords: Multiple sclerosis, Diffusion tensor imaging, Ising prior, Binary Markov random field

2.1 Introduction

Diffusion tensor imaging is a technique to quantify white matter pathways in the brain and spinal cord of living humans. In clinical applications, it opens the possibility to investigate the relationship between abnormal brain anatomy and neurological diseases (Ciccarelli et al., 2008). For example, several studies show that DTI can produce MRI indices in specific white matter tracts that may be associated with clinical disability in multiple sclerosis (MS), a disease that causes severe motor and cognitive deficits (Kern et al., 2010; Lin et al., 2008, 2005; Lowe et al., 2006; Ozturk et al., 2010).

These studies provide important insights into the organization of the brain and the effect of brain disorders. Results may be used as a tool for the diagnosis and management of patient care or as surrogate markers in future clinical trials, particularly if they are shown to be pharmacologically sensitive. However, some clinical researchers question the implications of these study results because the correlations between current MRI measures and clinical disability, although significant, have generally been low (Barkhof, 2002; Goodin, 2006). Such small correlations may be due to the intrinsic variability in the clinical expression of MS plaques in various anatomical locations.

Voxel-wise or mass-univariate regression, often referred to as the general linear model, is a standard technique for exploring the relationship between images and scalar measures of clinical disability. In this approach, we regress brain structure measurements on a disability score at each voxel separately Ashburner and Friston (2000),Smith et al. (2006)) to

produce a statistical parametric map (Friston et al., 1994). Such maps open the door to localizing the voxels that are significantly related to disability. However, mass-univariate estimation does not share information across neighboring voxels, and the resulting maps cannot be used to predict disability scores.

Multivariate or "decoding" models (e.g. Haxby et al., 2001; Haynes and Rees, 2006; Norman et al., 2006) seek to overcome these limitations. One such model that incorporates complex spatial structure is scalar-on-function regression (Goldsmith et al., 2011), in which the outcome is regressed on an entire one-dimensional white matter tract profile at once. This approach uses a weighted version of the tract profile, where the weights are estimated from the data. A useful by-product of the fitting algorithm is a tract-specific disability index, which is easy to understand and analyze. The method was developed for hundreds or thousands of locations along a neuronal tract, but it is not well suited for: 1) scaling up to tens or hundreds of thousands of locations; 2) modeling response surfaces that can be sparse and with abrupt edges; and 3) adapting to 3-D brain geometry, which contains complex manifold structures that are imperfectly observed.

In this paper we introduce a *scalar-on-image regression* method for studying the association between clinical measures and 3D brain maps. The method is computationally efficient, can be carried out over a large region of the brain, and can be adapted to highly irregular brain regions using a flexible spatial neighborhood definition. The term "scalar-on-image regression", analogous to the nomenclature of Reiss et al. (2011), refers to the fact that whereas the responses are scalars as in conventional regression, the predictors

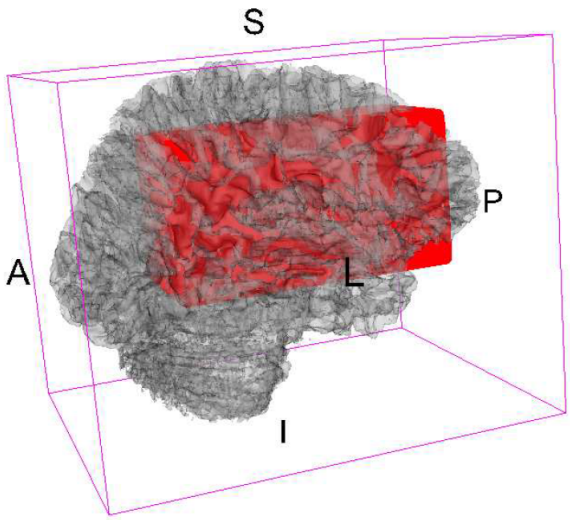

Figure 2.1: Red region contains the rectangular region we use as a predictor of cognitive function. Background 3D brain is rendered from a T1 template image.

are entire images. This method provides a *coefficient image* that describes the association between each voxel and the outcome, adjusting for all other voxels in the image. The proposed approach is Bayesian, adopting a novel prior that exploits both the presumed sparsity and the spatial smoothness of the coefficient image.

We apply our approach to data from a cross-sectional MRI study of multiple sclerosis (MS), and focus on studying the association between a clinical disability score and voxel-wise DTI indices in a large pre-specified region of the brain. More specifically, we use the PASAT score (Fischer et al., 1999) to measure cognitive disability and fractional anisotropy values to measure tissue viability. The region we consider is a $61 \times 125 \times 26$ collection of voxels including the corpus callosum (see Figure 2.1).

2.2 Methods

The core of our approach is to assume that there is an underlying unknown 0/1 map of voxels (associated/not associated with the outcome), and place an Ising prior on this latent binary image. Our model can be implemented through a single-site Gibbs sampler, where the computation time needed for each sweep over the image space is linear in the number of locations and does not depend on the number of nonzero coefficients.

We first introduce some notation. Assume the data for subject $i \in \{1, 2, ..., I\}$ are $\{y_i, X_i, Z_i\}$, where y_i is the scalar outcome (e.g. cognitive score), X_i is a vectorized image of the ith subject, and Z_i consists of other covariates (e.g. gender, age, etc.). In the MS example, every image X_i is a 3-dimensional array structure of dimension $L = L_1 L_2 L_3 = 61 \cdot 125 \cdot 26 = 198250$, though in general it can be an arbitrary 3-D manifold. We represent X_i as an $L \times 1$ dimensional vector , $(x_{i1}, x_{i2}, ..., x_{il}, ..., x_{iL})^T$, where x_{il} is an imaging measure, such as functional anisotropy, for subject i at voxel location l.

2.2.1 An ill-posed multiple linear regression

In essence, scalar-on-image regression is a multiple linear regression model, with the clinical outcome as the response and the whole image as the predictors:

$$\begin{aligned} y_i &= \alpha + Z_i^T \eta + X_i^T \beta + \varepsilon_i \\ &= \alpha + Z_i^T \eta + \sum_{l=1}^{L} x_{il}\beta_l + \varepsilon_i \end{aligned} \tag{2.1}$$

where $\beta = (\beta_1, \beta_2, ..., \beta_L)^T$ is a vector of coefficients for the image predictor X_i. In other words, each element, β_l is the coefficient for the image intensity x_{il} at voxel l. The parameter β_l can be interpreted as the change in y_i for each unit change in x_{il} adjusting for all other locations (i.e., $x_{il'}$ for all $l' \neq l$). The errors ε_i are independent and identically distributed normal random variables with mean 0 and variance σ_ε^2. See Figure 2.2 for an illustration.

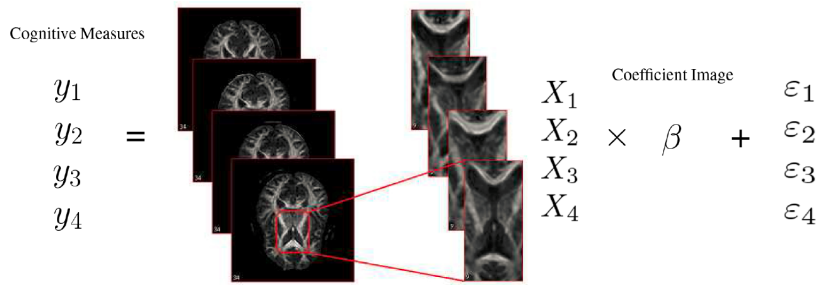

Figure 2.2: Illustration of the multiple linear regression model, with cognitive disability measure as the scalar response and fractional anisotropy maps as the image predictor.

When the intensities of all locations are mutually independent, solving this model will be equivalent to fitting separate linear regressions of y_i on x_{il} for each l. However, if the

voxel-level measurements are correlated, this multiple linear regression can in principle provide improved estimation by incorporating information across the brain as a whole.

We note that, whereas most predictive or "decoding" methods in neuroimaging (Haynes and Rees, 2006; Norman et al., 2006) have focused on pattern classification, Equation (2.1) models continuous outcomes (Cohen et al., 2011). The model can be extended to deal with classification problems, by assuming a discrete distribution for Y_i.

Unfortunately, fitting the multiple linear regression model (2.1) is an ill-posed problem. The dimension of X (here X is the collection of images across subjects, i.e. $X = (X_1, X_2, ..., X_I)^T$) is $I \times L$ and in most neuroimaging practice I (the number of subjects) is usually much smaller than L (the number of voxels), so that the least-squares solution is not unique. In order to get an estimate of the coefficient, certain dimension-reducing assumptions are needed to narrow the solution space. Our algorithm, presented below, narrows the solution space to a set of coefficient maps which are sparse and spatially continuous.

2.2.2 Fitting through penalization

A standard way to make the solution identifiable problem is penalized regression, wherein the usual least-squares criterion minimized in linear regression is replaced by a *penalized* least squares criterion, i.e., the estimate for model (2.1) is given by

$$\hat{\beta} = argmin_{\beta \in \mathbb{R}^L} \left\{ \sum_{i=1}^{I} (y_i - \alpha - Z_i^T \eta - X_i^T \beta)^2 + P(\beta) \right\}. \tag{2.2}$$

The penalty $P(\beta)$ is chosen to yield a solution to Equation (2.1) with desirable properties such as smoothness or sparsity.

Penalizing on β has a Bayesian interpretation: (2.2) is equivalent to enforcing a particular prior on the coefficients. Indeed, solving Equation (2.2) is statistically equivalent to the following model where the β coefficients are treated as random:

$$\begin{cases} y_i \sim N(\alpha + Z_i^T \eta + X_i^T \beta, \sigma_\varepsilon^2); \\ f(\beta) \propto \exp\left\{-P^{-1}(\beta)/2\right\}. \end{cases} \tag{2.3}$$

The solution $\hat{\beta}$ of model (2.2) equals the posterior mean $E(\beta|y)$ in Equation (2.3). The advantage of model (2.3) is that it provides a likelihood-based approach to fitting, which in turn allows inference on the model parameters. The second line of equation (2.3) means that β has a density function $f(\beta)$ proportional to $\exp\left\{-P^{-1}(\beta)/2\right\}$, where the normalizing constant is omitted. For specific forms of the penalty the distribution $\exp\left\{-P^{-1}(\beta)/2\right\}$ may not be integrable; in these situations model (2.3) may still provide reasonable results as in many cases the posterior distribution of $\beta|Y$ may still be proper even if the prior is not. However, in most cases the prior and the posterior distributions of β are proper.

One of the most popular penalties is the ridge regression or ℓ_2 penalty (Hoerl and Kennard, 1970) $P(\beta) = \lambda\beta^T\beta$, where λ is a scalar tuning parameter, with $\lambda = 0$ corresponding to no penalty and $\lambda = \infty$ corresponding to $\beta = 0$. Using (2.3), it follows that a ridge penalty is equivalent to assuming that the β parameters have an independent multi-

variate normal prior with constant variance. The lasso or ℓ_1 penalty (Bunea et al., 2011) $P(\beta) = \lambda \sum |\beta_l|$ is equivalent to a double exponential prior on β in (2.1). The same connection is also applied to elastic net penalty (Zou and Hastie, 2005; Carroll et al., 2009; Ryali et al., 2010; de Brecht and Yamagishi, 2012), whose corresponding prior is a mixture of normal and double-exponential.

Much recent work has been done to choose suitable spatial priors for neuroimaging data. For example, Penny et al. (2005) have proposed a fully Bayesian model with spatial priors defined over the regression coefficients of a general linear model, using Laplacian operators or Gaussian Markov Random Field. Flandin and Penny (2007) have proposed a Bayesian approach using a sparse spatial basis function priors. This model allows for spatial variations in intensity smoothness. As an alternative, Everitt and Bullmore (1999); Hartvig and Jensen (2000); Woolrich and Behrens (2006) model the spatial distribution of activation maps using mixture models.

2.2.3 Our proposal: Imposing an Ising prior on a latent 0/1 map

We are interested in priors that ensure that neighboring voxels have similar coefficient values and that non-zero coefficients form sparse patches in large areas of zero effects. Such local constraints are difficult to impose through the ridge or lasso penalties, as they assume that the β parameters are exchangeable and do not incorporate spatial dependence.

Thus, we focus on finding an appropriate prior distribution in the family of Markov random field spatial distributions. More precisely, we propose to use a neighborhood-based Ising prior (Cipra, 1987).

First, we introduce an L-dimensional binary random image γ such that $\beta_l = 0$ if $\gamma_l = 0$ and $\beta_l \neq 0$ if $\gamma_l = 1$; the binary map γ is a map that indicates which locations in the image coefficient are zero and do not impact the outcome. From an applied perspective, γ could be viewed as an unknown brain mask that defines regions of interest. Here we are interested in finding or estimating this mask. An Ising prior is used for γ, so that

$$p(\gamma = 1) = \phi(a, b) \exp\left[a\gamma + \sum_l \left\{ \sum_{l' \in \delta_l} bI(\gamma_l = \gamma_{l'}) \right\}\right]$$

where δ_l is the set of locations which are in the neighbourhood of location l and $\phi(a, b)$ is a normalizing constant. The parameters of the Ising distribution a and b control the overall sparsity and interaction between neighbouring points, respectively. Thus two assumptions are addressed: i) sparsity (controlled by a) — most voxels have coefficient $\beta_l = 0$ (no association with the measurement y_i); and ii) spatial contiguity (controlled by b) — a voxel is more likely to have a nonzero coefficient if its neighbors do. The parameters a and b could be allowed to vary spatially; for simplicity we assume that they are the same across locations.

Next, we assume that for those locations where the image is correlated with the outcome (i.e. $\gamma_l = 1$), β_l has a normal prior with an unknown variance σ^2_β. If $\sigma^2_\beta = +\infty$, then

no shrinkage will be placed on the estimated $\hat{\beta}$'s. We estimate σ^2_β by cross-validation, and in practice have found that a small σ^2_β achieves low prediction variance in noisy data sets. More precisely $[\beta_l|\gamma_l = 1, \beta_{-l}] \sim N(0, \sigma^2_\beta)$, which leads to the posterior conditional distribution,

$$\begin{aligned}[\beta_l \mid y, \gamma_l = 1, \beta_{-l}, \alpha, \eta] &\propto [y \mid \beta, \gamma_l = 1, \alpha, \eta][\beta_l \mid \gamma_l, \beta_{-l}] \\ &\sim N[\mu_l, \sigma^2_l],\end{aligned}$$

where $\mu_l = \sigma^2_l \left\{ \frac{1}{\sigma^2_\varepsilon}(y - \alpha - Z^T\eta - X^T_{-l}\beta_{-l})^T x_l \right\}, \sigma^2_l = \left(\frac{1}{\sigma^2_\varepsilon} x_l^T x_l + \frac{1}{\sigma^2_\beta} \right)^{-1}$ are the location-specific posterior mean and variance. Following the above equations, the location-specific posterior distribution comparing $(\gamma_l, \beta_l) = (0, 0)$ to $(1, \beta^*)$ is $p\{(\gamma_l = 1, \beta_l = \beta^*) \mid y, \beta_{-l}, \gamma_{-l}\} = \frac{1}{1+g}$ where

$$\begin{aligned} g &= \frac{p(y|\beta_l = 0, \beta_{-l})p(\gamma_l = 0 \mid \gamma_{-l})}{p(y \mid \beta_l = \beta^*, \beta_{-l})p(\beta_l = \beta^* \mid \gamma_l = 1)p(\gamma_l = 1 \mid \gamma_{-l})} \\ &= \exp\left[-\frac{1}{2\sigma^2_\varepsilon} \left\{ (Y - \alpha - Z^T\eta - X^T\beta^0)^T(Y - \alpha - Z^T\eta - X^T\beta^0) \right.\right. \\ &\quad -(Y - \alpha - Z^T\eta - X^T\beta^1)^T(Y - \alpha - Z^T\eta - X^T\beta^1)\big\} \\ &\quad \left.\left. +\frac{1}{2\sigma^2_l}(\beta^* - \mu_l)^2 - a + b\sum_{l' \in \delta_l}\{I(\gamma_l = 0) - I(\gamma = 1)\} \right] \sqrt{2\pi\sigma^2_l} \end{aligned}$$

Here, β^0 is the coefficient image corresponding to $(\gamma_l, \beta_l) = (0, 0)$ while β^1 is the coefficient image corresponding to $(\gamma_l, \beta_l) = (1, \beta^*)$, where β^* is sampled from the posterior distribution $[\beta_l|y, \gamma_l = 1, \beta_{-l}, \alpha, \eta]$.

Thus, at each image location the joint posterior distribution of the binary image and coefficient map is a Bernoulli choice that accounts for prior information through the Ising distribution as well as the relative impact of a zero and nonzero coefficient on the outcome likelihood.

2.2.4 Advantages of the Ising prior

The Ising prior has some important properties that are useful for conducting computations. Most importantly, the Ising distribution admits the single-site conditional distribution $p(\gamma_l = 1|\gamma_{-l}) = \frac{1}{1+g}$ where γ_{-l} is the vector of $\gamma_{l'}$ where $l' \neq l$ and

$$g = \frac{p(\gamma_l = 0|\gamma_{-l})}{p(\gamma_l = 1|\gamma_{-l})} = \exp\left[-a + b\sum_{l' \in \delta_l}\{I(\gamma_l = 0) - I(\gamma_l = 1)\}\right].$$

This indicates that the probability for voxel l to be predictive knowing the status of all other voxels in the brain depends only on the status of the voxels in a defined neighborhood of the voxel. (Here and below, we use "predictive" as shorthand for voxels with nonzero coefficients.) It also depends on the parameters a and b. This is intuitive and extremely helpful when one is interested in simulating from the posterior distribution of the latent 0/1 surface γ indicating whether a voxel is predictive or not. Indeed, instead of updating the entire image at once, one needs only update it one voxel at a time. This is why the algorithm is linear in the number of locations and remains relatively fast, even when the number of possible predictive locations is very large. Here we consider only contiguous,

cubic neighbourhoods, though other definitions are also possible.

2.2.5 Estimation

Our full model is

$$
\begin{aligned}
y_i &\sim N(\alpha + Z_i\eta + X_i\beta, \sigma_\varepsilon^2) \\
\beta_l &\sim \begin{cases} \delta(0), & \text{if } \gamma_l = 0 \\ N\left[0, \sigma_\beta^2\right] & \text{if } \gamma_l = 1 \end{cases} \\
\gamma_l &\sim \text{Bernoulli}[p_l] \\
p_l &\sim \text{Ising}[a, b]
\end{aligned}
$$

where $\delta(0)$ is a point-mass at zero. The Ising prior controls the number of nonzero coefficients and favours contiguity of localized effects. The Bernoulli choice between zero and nonzero coefficients at each location depends the posterior probability whether a voxel is predictive or no. Goldsmith et al. (2013) imposed a conditional autoregressive (CAR) prior on β_l whose indicator variable γ_l equals 1 and used a much smaller number of predictor voxels (30K). Here we use an exchangeable prior on the size of effects at those locations that are found to be associated with the outcome.

We implement a single-site Gibbs sampler to generate iterates from the posterior distribution of (γ, β). At the tth step, we proceed through the following steps for each location

$l \in \{1, 2, ..., L\}$:

1. Calculate μ_l, σ_l^2 from $\beta_{-l}^{(t)}$
2. Generate $\beta_l^1 \sim N(\mu_l, \sigma_l^2)$
3. Calculate the posterior probability g from $\beta_{-l}^{(t)}$ and β^1
4. Generate $\gamma_l^{(t+1)} \sim$Bern(g)
5. If $\gamma_l^{(t+1)} = 1$, $\beta_l^{(t+1)} = \beta_l^1$; otherwise $\beta_l^{(t+1)} = 0$.

2.2.6 Tuning parameters

The parameters a, b, σ_ε^2 and σ_β^2 control the estimation of the coefficient map and are referred to as tuning parameters. Here σ_ε^2 determines the impact of the change in the outcome likelihood on the overall probability whether a voxel is predictive or not. Similarly, in the posterior distribution of predictive regression coefficients, the parameter σ_β^2 is important in determining the posterior mean and variance. Finally, a controls the overall sparsity, while b determines the overall degree of smoothness among the γ parameters.

To select these tuning parameters we use five-fold cross validation. Our data are divided into five randomly selected groups. Each time, we obtain the training model from four groups and calculate the prediction error from the rest group. The procedure is repeated 5 times and the average of the prediction errors is calculated; the tuning parameters estimators (a, b, σ_ε^2, σ_β^2) minimize this average prediction error.

The model provides an excellent exploratory and sensitivity analysis tool where results can be inspected by simply changing the tuning parameters. We find this multi-resolution approach to be very helpful in the context when one is interested in further exploring results beyond simply using the cross validated values. Such an exploratory analysis could be based on modifications of the estimated tuning parameters.

2.2.7 DTI study

As discussed in the Introduction, our motivating application is to investigate the relationship between cognitive disability in multiple sclerosis patients and their diffusion tensor images. Multiple sclerosis is an immune-mediated disease that affects the brain and spinal cord (central nervous system). It results in damage to the myelin sheath, the protective covering that surrounds axons in white matter. Damage caused by MS can disrupt the transmission of signals in affected tracts.

Study participants with MS were recruited from an outpatient neurology clinic and healthy volunteers from the community. All disability scores were measured within 30 days of the MRI scan. Prior to MRI scanning and disability testing, all participants gave signed, informed consent. All procedures were approved by the NIH review board (Goldsmith et al., 2011).

We used the Paced Auditory Serial Addition Test (PASAT) as a proxy measurement for cognitive disability. This score assesses mental capacity, rate of information processing and

sustained and divided attention. The range of PASAT is from 0 to 60, with higher scores indicating better cognition ability (Fischer et al., 1999).

All DTI scans were performed on a 3T scanner (Intera; Philips, Best, The Netherlands) over a 4.6 year period, using the body coil for transmission and either a 6-channel head coil or the 8 head elements of a 16-channel neurovascular coil for reception (both coils are made by Philips). Each session included two sequential DTI scans using a conventional spin-echo sequence and a single-shot EPI readout. Whole-brain data were acquired in nominal 2.2 mm isotropic voxels and with the following parameters: TE, 69 ms; TR, automatically calculated ("shortest"); slices, 60 or 70; parallel imaging factor, 2.5; non-collinear diffusion directions, 32 (Philips "overplus high" scheme); high b-value, 700 s/mm^2; low b-value ("b_0"), approximately 33 s/mm^2; repetitions, 2; reconstructed in-plane resolution, 0.82×0.82 mm. A 3D gradient-echo magnetization-transfer sequence was performed with segmented EPI readout (nominal acquired resolution, $1.5 \times 1.5 \times 2.2$ mm; TE, 15 ms; TR, 64 ms; parallel imaging factor, 2; EPI factor, 7; magnetization-transfer pulse, sinc-shaped, 1.5 kHz off-resonance; repetitions, 3); the resulting images were rigidly registered to the DTI scan before calculation of MTR maps (defined as 1 minus the voxel-wise ratio of data from this sequence to those obtained using the same sequence without the magnetization-transfer pulse). Prior to analysis, data were adjusted to account for changes in average tract-specific MRI indices that resulted from scanner upgrades,by a procedure previously described by Harrison et al. (2011).

Here we focus on fractional anisotropy (FA) (Cercignani et al., 2001; Hasan et al.,

2005). The diffusion-weighted scans were processed using CATNAP (Landman et al., 2007) to create maps of FA. The whole-brain FAs were calculated by slice-wise averaging of all diffusion-weighted images, removal of the low-intensity voxels that are characteristic of extracerebral tissues on these images, and final removal of voxels with MD > 1.7 μm^2/ms to exclude cerebrospinal fluid (Ozturk et al., 2010). The resulting brain mask was applied to all DTI maps.

In summary, our study consists of data from 135 MS patients. Each has one PASAT score and one FA image with dimension $61 \times 125 \times 26$, registered to ensure major structures (e.g., the corpus callosum) are aligned across subjects.

No. of participants (% women)	135 (35%)
Mean age, years (SD; range)	44 (12; 20-69)
Mean PASAT (SD)	44 (13)

Table 2.1: MS patient characteristics. Disability data were obtained within 30 days of the MRI scan.

2.3 Results

After choosing the tuning parameters $a, b, \sigma_\varepsilon^2, \sigma_\beta^2$ by cross-validation, we run the image regression model through the Gibbs sampling algorithm. We use a chain of length 500 and discard the first 100 samples as burn-in. All the regression coefficients and latent binary indicators are initialized at 0.

Figure 2.3 shows the overall distribution of estimated regression coefficients in β and

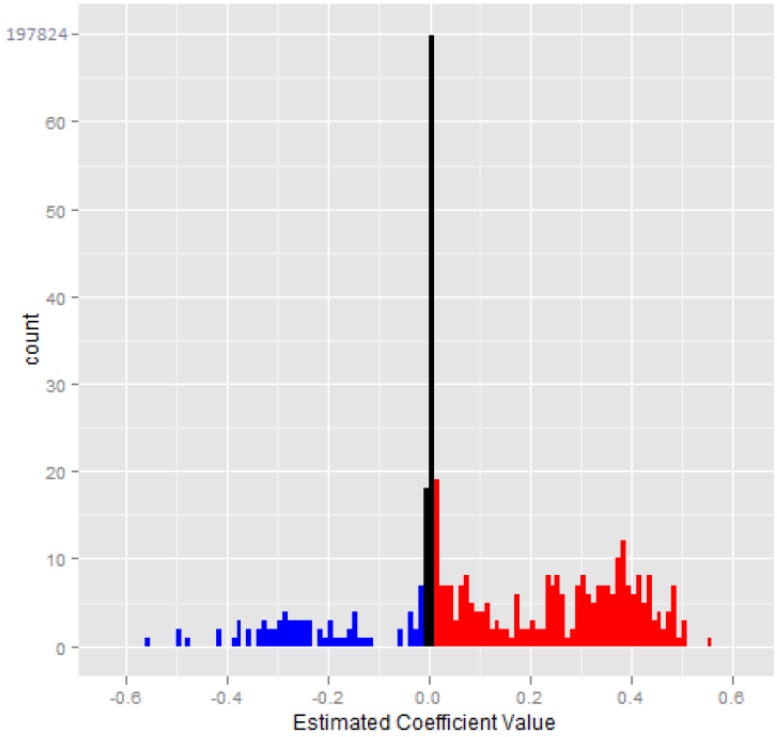

Figure 2.3: Histogram of the estimated coefficients with tuning parameter values $a = -3, b = 6, \sigma_\varepsilon^2 = 1.22, \sigma_\beta^2 = 0.05$, which were chosen by five-fold cross-validation. The middle bar refers to coefficients whose magnitude lies below a threshold of 0.01. The blue bars denote the coefficients which are less than -0.01 while the red ones denote the coefficients that are larger than 0.01.

Figure 2.4 shows the estimated coefficients overlaid on an anatomical reference from Slice 7 to Slice 22. The first thing to notice is that most of coefficients are zero (426 of the 197842 voxels had $\beta_l \neq 0$), due to sparsity-inducing effect of the Ising prior. Figure 2.3 indicates that the number of coefficients with positive coefficient estimates (red lines) is larger than the number of negative estimates (blue lines). Thus, in most of the predictive voxels, lower FA values correspond to lower PASAT scores. This agrees with the scientific hypothesis that degradation of white matter is associated with diminished cognitive ability. Moreover,

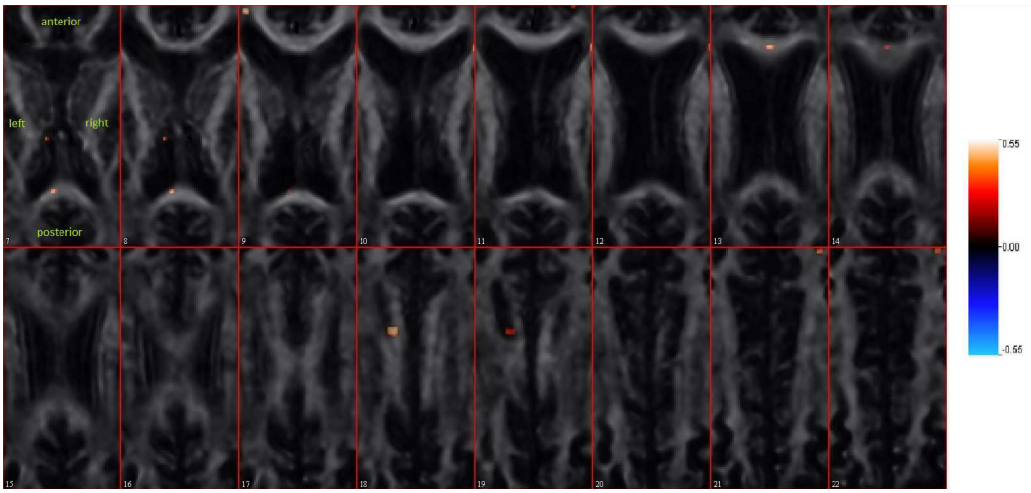

Figure 2.4: The estimated coefficient images from Slice 7 to Slice 22. The estimation is overlaid on one single subject's FA scan image for anatomical reference. The tuning parameters are selected via cross validation, $a = -3, b = 6, \sigma_\varepsilon^2 = 1.22, \sigma_\beta^2 = 0.05$. Positive coefficients are shown in red, while blue denotes negative coefficients. The estimated mean square prediction error is 146.43 and the proportion of variance explained for predicted data is 22.00%.

in Figure 2.4 the "visible" predictive regions, though extremely sparse, are located in the corpus callosum — the largest white matter structure in the brain, which has been related to cognitive ability (Ozturk et al., 2010).

Up to this point, we have used cross-validation to select tuning parameters and have provided estimated coefficient images that satisfy the sparsity and spatial continuity assumptions. However, one might be interested in exploring results as one moves away from the optimal cross-validated values of parameters. In fact, only 426 out of 198250 voxels have nonzero coefficients, probably because data are noisy and cross validation heavily penalizes coefficients. This helps prediction performance but may be too restrictive when one is interested in exploratory data analysis and hypothesis generation. For exploratory

purposes, one may be interested in obtaining less conservative coefficient images.

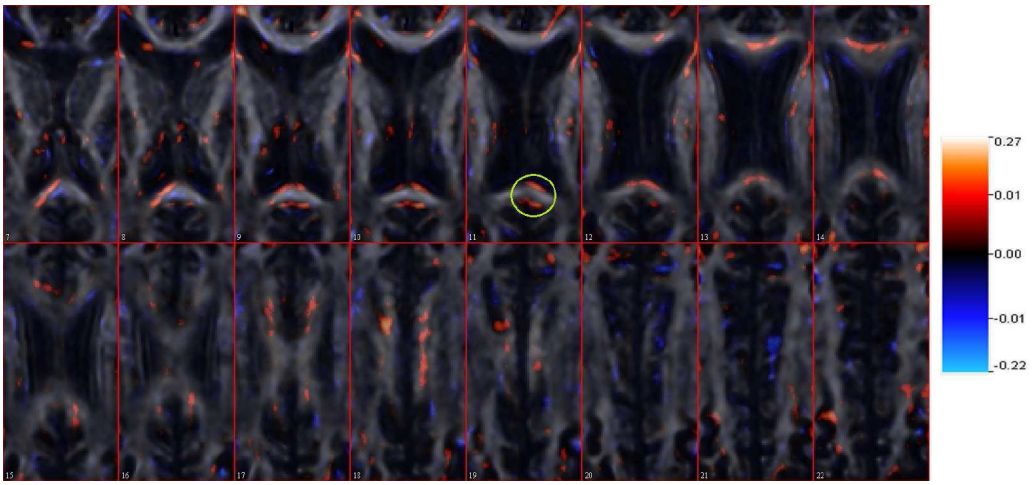

Figure 2.5: The estimated coefficient images from Slice 7 to Slice 22 using tuning parameters $a = -2, b = 0.5, \sigma_\varepsilon^2 = 1, \sigma_\beta^2 = 0.03$. The estimation is overlaid on one single subject's FA scan image for anatomical reference. The estimated mean square prediction error is 150.25 and the proportion of variance explained for predicted data is 20.34%.

Figures 2.5 and 2.6 present the coefficient images that result from two other combinations of tuning parameters. Starting from the cross-validation setting (Figure 2.4), we select the tuning parameters (i.e. increase a, decrease b, decrease σ_β^2 and decrease σ_ε^2) so that the estimation becomes less conservative. While a higher number of predictive regions are revealed from Figure 2.4 to Figure 2.6, the prediction power of the corresponding models is decreased. Table 2.2 shows the estimated mean of squares of prediction errors and the proportion of variance explained for predicted data corresponding to Figures 2.4 through 2.6. From this we can see that a range of coefficient images can provide similar results in terms of prediction power, and the choice of final model depends on both prediction accuracy and interpretation of the final result.

Figure 2.7 presents profile cross-validation plots to investigate the effect of each tuning

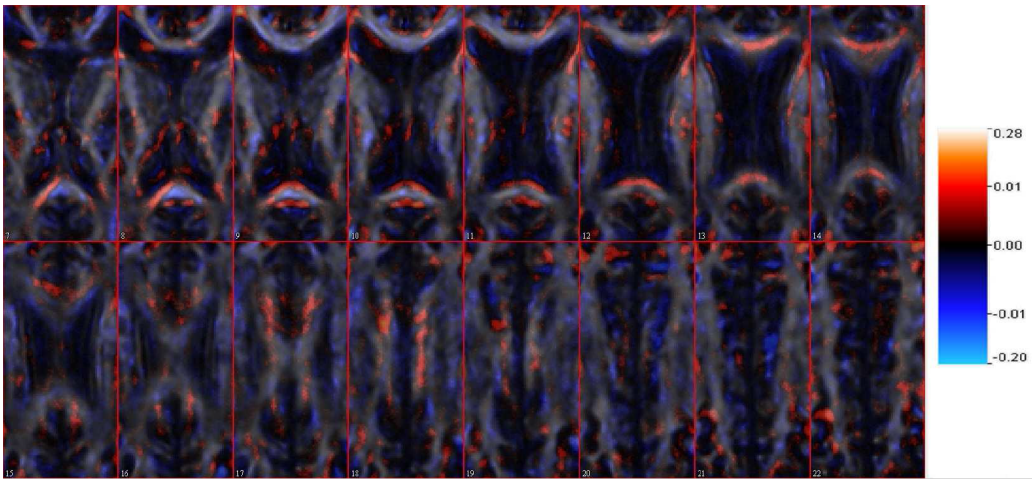

Figure 2.6: The estimated coefficient images from Slice 7 to Slice 22 using tuning parameters $a = -1, b = 0.5, \sigma_\varepsilon^2 = 0.775, \sigma_\beta^2 = 0.05$. The estimation is overlaid on one single subject's FA scan image for anatomical reference. The estimated mean square prediction error is 164.6 and the proportion of variance explained for predicted data is 19.22%.

Figure	σ_ε^2	a	b	σ_β^2	MSE	PVE
2.4	1.22	-3	6	0.05	146.43054	0.22000
2.5	1	-2	0.5	0.03	150.25746	0.20341
2.6	0.775	-1	0.5	0.05	164.60563	0.19220

Table 2.2: Prediction performance of the tuning parameter combinations for Figures 2.4–2.6 . The last two columns refer to mean square prediction error and proportion of variance explained.

parameter on the performance of the model. In each of the four panels, three of the tuning parameters are fixed at the optimal values chosen by cross-validation while the remaining tuning parameter varies in the x-axis. As shown in the figure, tuning parameters a, b have relatively small influence on the prediction performance. Thus, the empirical choice of those two parameters can be more flexible. Also, as σ_β^2 increases, the proportion of variance explained in left-out data drops dramatically, which indicates the shrinkage of β's is necessary.

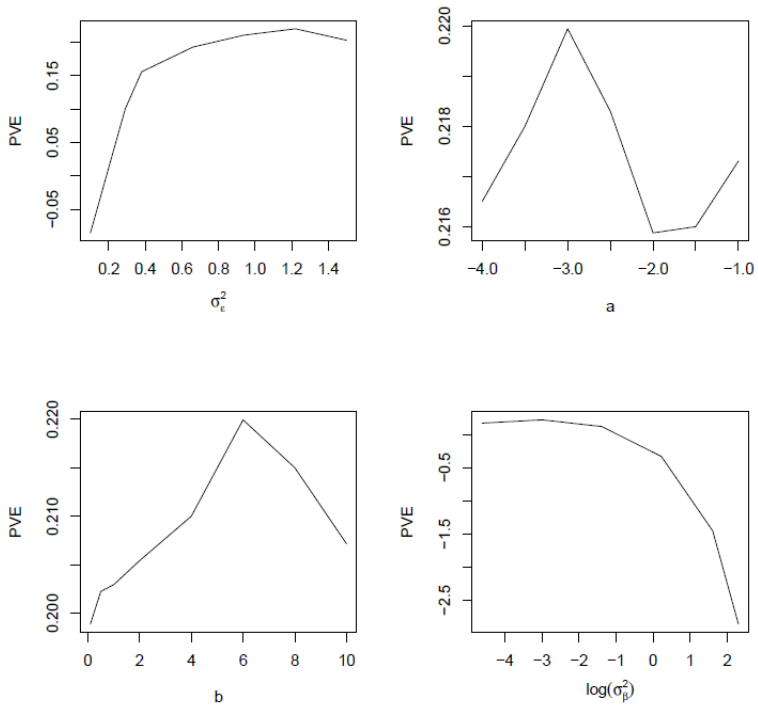

Figure 2.7: Profile cross validation plot: in each panels, three of the tuning parameters are fixed at the values chosen by cross-validation while the remaining tuning parameter varies in the x-axis. The y-axis is the proportion of variance explained in the left-out data.

Balancing between prediction accuracy and result interpretation, in Figure 2.5, we choose $a = -2, b = 0.5, \sigma^2_\varepsilon = 1, \sigma^2_\beta = 0.03$ as tuning parameters. In the estimated coefficient image, 41455 out of 198250 voxels have nonzero coefficients. Although significant effects tend to be located in the anterior region in the left side comparing to a scattered pattern in the right side, the coefficients on both sides maintain spatial contiguity. Most positive coefficients are located in the corpus callosum, which indicates that cognitive ability may be positively associated with integrity of the white matter in that region. We also found negative coefficients outside corpus callosum (e.g., in the posterior right region of Slice 21). This might be due to the undersmoothed estimation caused by a small b value,

though further investigation is necessary. Several predictive regions (marked with a green circle) are located on the edge of the white matter, possibly due to registration error. In the future, we may apply the registration technique adopted in Tract-Based Spatial Statistics (TBSS) (Smith et al., 2006). A scalar-on-image regression model can then be fitted in the "mean FA skeleton" space.

For comparison, we performed voxel-wise regressions, with PASAT score regressed on the FA values for each voxel in 198250 separate linear regressions. (Note that in standard mass-univariate regression, the roles of PASAT and FA would be reversed.) In Figure 2.8, we plot the uncorrected p-values of the slope coefficients from Slice 7 to Slice 22. Most voxels with small p-values are located in the corpus callosum. Moreover, the regions with small p-values in the voxel-wise regression tend to have large coefficients in our scalar-on-image regression.

Comparing the results in Figure 2.5 with Figure 2.8, the voxels with small p-values in Figure 2.8 spread symmetrically to the left and right part of the brain while our method shows an asymmetric pattern of predictive coefficients. For example, in Slice 18 there are predictive coefficients located in the left anterior region while in the right part, the significant coefficients are more evenly distributed across the corpus callosum. This difference is caused by different assumptions imposed by the two models. In the scalar-on-image regression model, we place sparsity and spatial contiguity assumptions on the coefficient image, incorporating spatial structure during the estimation step. For voxel-wise regression, homotopic correlations are not adjusted for in the estimation, resulting in a symmetric p-value

map.

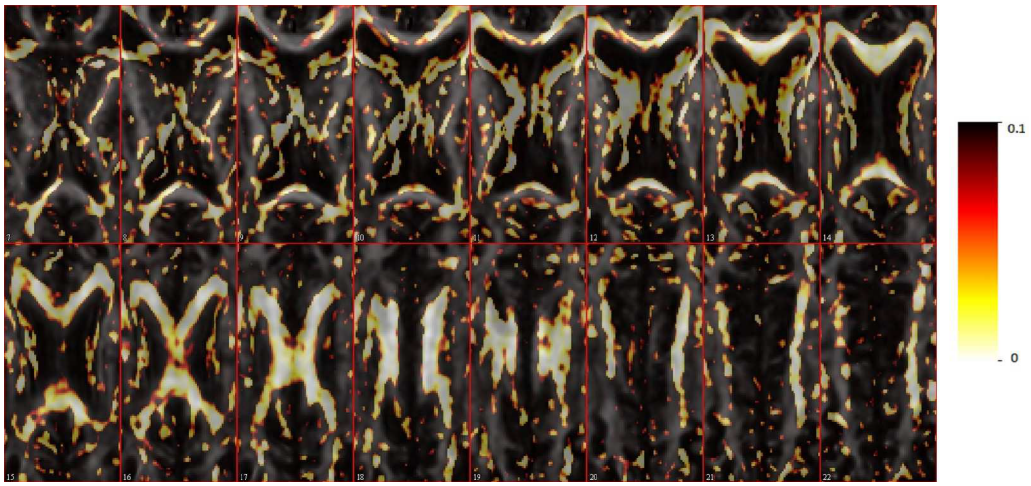

Figure 2.8: p-value map for voxel-wise linear regression fitting from Slice 7 to Slice 22.

2.4 Discussion

We have proposed a novel linear regression approach for analyzing the relationship between cognitive disability and white matter microstructure in three-dimensional images. Noting the connection between penalized linear regression and Bayesian modeling, we proposed a Bayesian regression model with a latent binary indicator. We take advantage of an Ising prior to impose the assumption of sparsity and spatial continuity in the analysis. A distinctive merit of our method is that the regression model can be established on any manifold. By contrast, most scalar-on-image regression approaches (e.g. Reiss and Ogden, 2010; Reiss et al., 2015) require a regular grid. For simplicity, in our application we focused on a rectangular region, but the method is easy to extend to any irregularly shaped region,

including the entire brain.

We applied our model to a multiple sclerosis study. The results show most of the predictive regions are located at the corpus callosum, as expected from existing work (Barnea-Goraly et al., 2004; Keller et al., 2007).

There are a few limitations in the presented methodology. First, if we choose the hyper-parameters via cross-validation, the computation time is high; this can be partially alleviated by parallel computing. Alternatively, a pilot cross-validation study can be done on part of the image region and the estimated parameters can then be applied to the entire image. Moreover, our approach is a hybrid between a Bayesian and a frequentist approach, where the hyper-parameter and coefficient estimation proceeds in parallel. A fully Bayesian approach might provide a more integrated and philosophically satisfying alternative. Third, we emphasize the sparsity of the coefficient image. In some analyses, one may be interested in borrowing strength from the immediate neighbours, as done via the CAR prior in Goldsmith et al. (2013). Lastly, the predictive regions in our application did not localize particularly well in the white matter. This may be caused by registration error, which can be improved by TBSS in the previous session. Also, our current model only incorporates the neighbourhood information and emphasize on the sparsity of the predictive regions. We can also consider putting extra constraints on the coefficients to force the regions in white matter to have higher probabilities to be predictive.

Avenues for further work include the following. (1) Instead of a continuous response variable, we can extend our model to cope with categorical variable for classification prob-

lem. A Metropolis-Hastings algorithm will be implemented to sample from the conditional posterior distribution during Gibbs sampling. (2) We will develop inferential tools for statistical testing for image regression. As an analogy to the confidence band in frequentist inference (Reiss and Ogden, 2010), we can construct the credible interval (or Bayesian posterior interval) from the Gibbs samples. (3) In terms of application, we may consider extending the analysis to the entire brain or using other DTI measures. It is also simple to extend our method to single-subject fMRI data. For multiple-subject fMRI data, one possible solution is to incorporate a spatio-temporal process into the prior of the scalar-on-image regression model (Woolrich et al., 2004).

Chapter 3

Two-way PCA for matrix-variate data

Abstract

Many modern neuroimaging studies acquire large spatial images of the brain observed sequentially over time. Such data are often stored in the forms of matrices. To model these matrix-variate data we introduce a class of separable processes using explicit latent process modeling. To account for the size and two-way structure of the data, we extend principal component analysis to achieve dimensionality reduction at the individual level. We introduce necessary identifiability conditions for each model and develop scalable estimation procedures. The method is motivated by and applied to a functional magnetic resonance imaging study designed to analyze the relationship between pain and brain activity.

***keywords*: fMRI, latent process modeling, matrix-variate, principal component analysis, separability**

3.1 Introduction

Blood-oxygen-level dependent functional magnetic resonance imaging (BOLD fMRI) measures brain activity by detecting changes in blood oxygenation and blood flow related to neuronal activity (Huettel et al., 2004), thereby providing researchers with means to study human brain function *in vivo*. During a standard fMRI experiment, several hundred brain images (each comprising measurements at roughly 100,000 volume units, or voxels) are acquired while the subject performs a sequence of tasks. Changes in the measured signal between images are then used to make inferences regarding possible task-related activations in the brain. Over the past two decades, fMRI has been used to successfully localize regions of the brain activated by a task, determine distributed networks that correspond to brain function, and make predictions about psychological or disease states (Lindquist, 2008). Standard fMRI data exhibit a complicated two-way (spatial and temporal) structure with a relatively weak signal. Hence, data are not only massive in scale but also highly complex.

The methodological developments presented here were motivated by an fMRI study of thermal pain performed on 20 subjects (Lindquist, 2012). Each subject was scanned while subjected to a series of pain trials, consisting of thermal stimulations to the left volar forearm. The number of trials for each subject varied from 45 to 52, with a total of 940 trials for the 20 subjects. For each trial, one of two levels of thermal simulation was randomly assigned: temperature was calibrated to be either painful (hot) or non-painful (warm) using a pain calibration task performed prior to scanning. Each trial lasted 46 seconds. Following

an 18-second interval consisting of thermal stimulation, a fixation cross was presented for a 14-second interval at the end of which the words "How painful?" appeared on a screen. After a few seconds of contemplation, the participant rated the overall pain intensity between 100 and 550, where larger values indicate more pain. Figure 3.1 illustrates the stimulation design for a single trial. Prior to analysis, data were extracted from 21 different classic pain-responsive brain regions. For each 46 second trial, measurements were made every 2 seconds resulting in 23 sampling points. Thus, the data we consider in this study consist of 21 locations, observed over 940 trials, each consisting of a time series of length 23 (corresponding to 46 seconds of brain activation). We denote the observed fMRI data for each trial $i \in \{1, 2, ..., 940\}$ as $Y_i(s, t)$ where $s \in \{1, 2, ..., 21\}$ and $t \in \{1, 2, ..., 23\}$ are the indexes for the spatial and temporal domains respectively. These data have a natural two-way space-time structure. In Figure 3.2 we display the fMRI data for two randomly selected trials from the first three subjects. In each panel, the x-axis represents time from 0 (start of thermal stimulation) to 46 seconds (end of pain rating) and the y-axis represents the BOLD fMRI signal from the 21 different classic pain-responsive brain regions indexed from 1 to 21 (see Table S.1 in Supplementary Materials). There was no known a-priori spatial correlation between these regions. The fMRI intensity is color-coded from dark red (low intensity values) to light yellow (high intensity values). One of the primary goals of this study is to detect fMRI patterns that may correspond to pain stimuli.

One possible first step, which is the focus of this paper, is to do principal component analysis (PCA) for matrix-type data. The reason we focus on PCA is that it is a first-

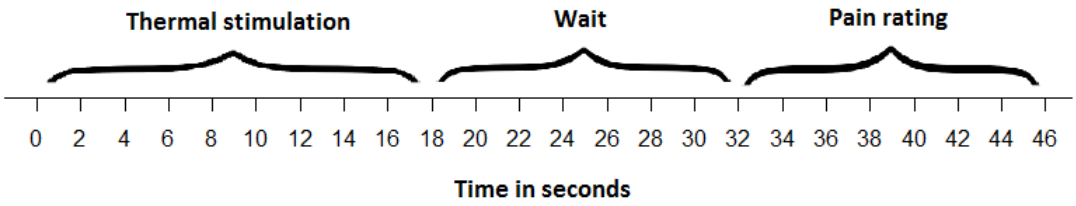

Figure 3.1: The stimulation design for a single trial during a 46-second time course.

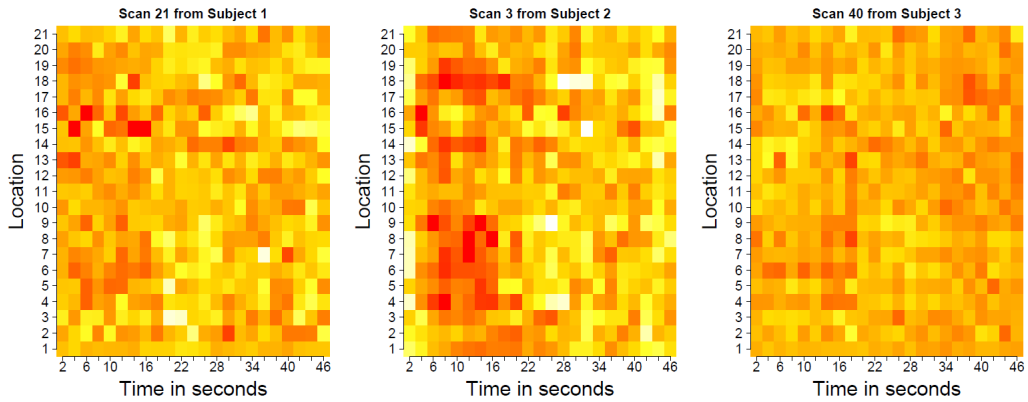

Figure 3.2: fMRI intensity measurements over 21 locations $\times$ 23 time points over two subjects.

line approach for pattern capturing and dimensionality reduction and is also widely used in fMRI studies. Ror example, when performing independent component analysis (ICA), one often starts with a time-domain PCA (PCA of the time series) as a preprocessing step to reduce the dimensionality and mitigate the effects of noise (Hyvärinen and Oja, 2000; Calhoun et al., 2001). Moreover, several authors (e.g. Viviani et al., 2005; Lindquist, 2008) have advocated the use of PCA as a data-driven approach to characterize patterns in the data.

There are three major approaches for using PCA directly in the analysis of fMRI data. In the first and second approaches, one could simply stack scans along one domain (time or

location) and conduct a temporal or spatial PCA along the other domain. For example, spatial stacking could be achieved by simply vectorizing the data for every location. For region s, this would result in the vector $[Y_1(s,1), Y_1(s,2), ..., Y_1(s,23), Y_2(s,1), ..., Y_{940}(s,23)]$. This vector can then be row stacked resulting in a 21 by 21620 dimensional matrix. A PCA on the columns of this matrix is typically referred to as a spatial PCA and produces eigenimages. A similar stacking can be done to produce temporal PCA. The third approach is the population value decomposition (PVD) proposed by Caffo et al. (2010) and Crainiceanu et al. (2011). PVD uses a two-stage singular value decomposition (SVD) to extract population-level principal components along each dimension. The first-stage SVD is implemented on each $Y_i(s,t)$ and the second stage SVD is implemented on derived eigenvectors from the previous step.

What we propose here is related to these approaches, but is fundamentally different because we explicitly model row space and column space separately. More specifically, we incorporate the two-way structure of fMRI data into PCA by explicit modeling using latent processes under specific separability assumptions. The main contributions of this paper are as follows. First, by introducing explicit separability assumptions, the dimensionality of the covariance matrix will be substantially reduced. Indeed, PCA on concatenated data would require diagonalizing $S^2 \times T^2$-dimensional matrices whereas PCA of separable process would be of dimension $S^2 + T^2$, where S and T are the numbers of rows and columns for each observed matrix, respectively. In our fMRI study, $S^2 \times T^2 = 233,289$ and $S^2 + T^2 = 970$. Second, the induced two-way PCA will provide an explicit decomposition of the data

according to the separability assumptions. This, coupled with a linear mixed-effect model (LMM) will allow for future model building for longitudinal or hierarchical data. Third, three different types of separable models (additive, multiplicative and hybrid) are proposed, which will be useful under different scenarios. Fourth, fast method-of-moments based algorithms are derived and the identifiability conditions for each model will be carefully studied.

Two-way matrix modeling is a highly active research area. Recently, Allen et al. (2014) introduced a two-way SVD for matrices to account for possibly correlated residuals. The correlation structure of the residuals was accounted for by using a separability assumption on the residuals. In contrast, we impose separability on the signal part of the model. This is achieved by explicit modeling of the scan-specific latent processes. This allows our model to be easily generalized to more complex data that may include multi-level and longitudinal designs as well as non-continuous outcomes. There are many developments in the area of regularized PCA for matrix data and two-way functional data (Allen, 2013b; Huang et al., 2008; Witten et al., 2009; Lee et al., 2010; Huang et al., 2009). The focus of our paper is not on regularization, although regularization could be used before our methods are implemented; see, for example, the sandwich smoother for very large matrices (Xiao et al., 2013). Another area of research is concerned with scalar-on-matrix regression (Zhou and Li, 2014; Zhou et al., 2013). These approaches are different from ours because they focus on regression, whereas we are concerned with discovery of patterns in population level of matrices independent of the outcomes. Spencer et al. (2001) and Dien et al. (2003)

proposed a two-stage PCA, which perform a spatial PCA and a temporal PCA sequentially while our method estimates the principal components from the two domains simultaneously under explicit model separability assumptions.

The rest of the paper is organized as follows. In Section 3.2, we introduce three models based on different separability assumptions. We discuss their corresponding identifiability conditions and derive their covariance structures. In Section 3.3, we propose the estimation algorithm for the two-way PCA. In Section 3.4, we extend our method to data with white-noise. Section 3.5 provides a simulation study while Section 3.6 provides extensive results for the analysis on the fMRI pain study. We conclude the paper with a discussion of potential future research in Section 3.7. All technical proofs are delegated to the Supplementary Materials.

3.2 Separable models for two-way matrix data

In this paper, "separability" refers to the property that the variability of observed matrices can be divided into row-specific and column-specific components. We introduce three different types of separability: additive, multiplicative and hybrid.

3.2.1 Separable additive model

The separable additive model is

$$Y_i(s,t) = \mu(s,t) + C_i + U_i(s) + V_i(t), \tag{3.1}$$

where letters s and t denote the indices for rows and columns in the matrix. In model (3.1), the observed matrix outcome $Y_i(s,t)$ is decomposed as a linear sum of a deterministic mean matrix, $\mu(s,t)$, a scan-specific and row-column invariant random deviation C_i, a row-specific random variable $U_i(s)$ and a column-specific random variable $V_i(t)$. The random variables C_i are assumed to be i.i.d. (independently and identically distributed) with mean zero and variance V_C while $U_i(s)$ and $V_i(t)$ are assumed to be i.i.d. with mean zero and covariances $\Sigma_U(s,s') = E\{U_i(s)U_i(s')\}$ and $\Sigma_V(t,t') = E\{V_i(t)V_i(t')\}$, respectively.

The additive representation of $Y_i(s,t)$ is not unique: adding a constant to $U_i(s)$ and subtracting the same constant from $V_i(t)$ will not change their sum. The following result provides necessary identifiability conditions for model (3.1).

Theorem 1. *Consider the separable additive model* (3.1) *and assume that the following conditions are satisfied for all s and t.*

A.1 C_i, $U_i(s)$ and $V_i(t)$ have finite second moments.

A.2 $E\{C_i\} = E\{U_i(s)\} = E\{V_i(t)\} = 0$.

A.3 (Separability) $C_i, U_i(s)$ and $V_i(t)$ are mutually uncorrelated.

A.4 $\sum_{s_1=1}^{S}\sum_{s_2=1}^{S} Cov\{U_i(s_1), U_i(s_2)\} = 0$ *and* $\sum_{t_1=1}^{T}\sum_{t_2=1}^{T} Cov\{V_i(t_1), V_i(t_2)\} = 0$.

Then the second order moments of C_i, $U_i(s)$ *and* $V_i(t)$ *are identifiable. Moreover, if* $[C_i, U_i(s), V_i(t)]$ *is jointly normal, the distributions of* C_i, $U_i(s)$ *and* $V_i(t)$ *are also identifiable.*

Supplementary Materials S.2 provides the proof of this theorem. While assumptions A.1 through A.3 are relatively standard, assumption A.4 is less so. The intuition for assumption A.4 is that all row- and column-invariant variability in the data is captured by the random effect C_i. Assumption A.4 can be shown to imply that all the eigenvectors of $U_i(s)$ and $V_i(t)$ are orthogonal to the constant.

3.2.2 Separable multiplicative model

A different way to introduce separability is via the multiplicative model

$$Y_i(s,t) = \mu(s,t) + U_i(s)V_i(t). \tag{3.2}$$

Assuming $U_i(s)$ and $V_i(t)$ are uncorrelated, the covariance matrix of Y evaluated at all s's and t's can be written as the Kronecker product $\mathbf{\Sigma}_Y = \mathbf{\Sigma}_U \otimes \mathbf{\Sigma}_V$, where $\mathbf{\Sigma}_U$ and $\mathbf{\Sigma}_V$ are the covariance matrices for $U(s)$ and $V(t)$.

Kronecker product covariance structures have been used extensively for dimension reduction when modeling spatio-temporal processes (Cressie, 1993; Genton, 2007). How-

ever, the meaning of "separability" in this section is fundamentally different from the one conventionally used in spatio-temporal statistics. First, our definition of separability is explicit and provides a platform for generalization that can be applied to the additive and hybrid separability. Second, our approaches apply to *samples* of spatio-temporal matrices, whereas the current spatio-temporal process literature is typically applied to one realization of the process. Thus, identifiability problems and solutions are quite different. Third, in the spatio-temporal statistical literature, separability is used directly on the covariance matrix, while in our case the separability is induced by modeling the latent row-specific and column-specific processes. Our model (3.2) implies the separable covariance structure, though a Kronecker-product covariance structure for the observed process does not imply the modeling structure in model (3.2).

Multiplicative separable models are non-identifiable as $U_i(s)V_i(t) = \{aU_i(s)\}\{\frac{1}{a}V_i(t)\}$ for any positive constant a. Therefore, in addition to the standard assumptions we add the following constraint on the latent spatial process $U_i(s)$: Var$\{U_i(0)\} = 1$. In Supplementary Materials S.3, we present the complete list of identifiability conditions in Theorem 2 and its proof.

3.2.3 Separable hybrid model

Our third model is a combination of the separable additive and multiplicative models:

$$Y_i(s,t) = \mu(s,t) + C_i + U_{1i}(s) + V_{1i}(t) + U_{2i}(s)V_{2i}(t). \tag{3.3}$$

As in models (3.1) and (3.2), C_i is the row-column invariant component, $U_{i1}(s)$ is the column-invariant component, $V_{i1}(t)$ is the row-invariant component, and $U_{i2}(s)V_{i2}(t)$ is the first-order interaction between row and column spaces. To ensure identifiability, we assume that C_i is uncorrelated with $U_{1i}(s), V_{1i}(t), U_{2i}(s)$ and $V_{2i}(t)$ and all row-specific processes are uncorrelated with the column-specific processes. However, zero correlation among the row-specific as well as the column-specific processes is not required. The detailed identifiability conditions (Theorem 3) can be found in Supplementary Materials S.4.

Model (3.3) is formally similar to the additive main effects and multiplicative interactions (AMMI) model (Gauch, 1988), but there are several differences: (a) As in traditional spatio-temporal statistics, the AMMI model is often used for a single matrix whereas our model focuses on population level analysis. (b) The AMMI model aims to estimate the separable processes themselves while our model is focusing on the covariance structure. Introducing latent separable processes is a middle step between real data and PCA, which makes PCA result more interpretable. c) The AMMI model was first introduced as a fixed effect model while ours treats the latent processes as random effects. Smith et al. (2001) extended the AMMI model to have a random multiplicative interaction. However, the co-

variance structure induced is still restricted.

We summarize our three separable models in Table 3.1. In the last column of the table, the total variability of the original process is decomposed into component-specific variations, where J_T and J_S are matrices of dimension $T \times T$ and $S \times S$ with all elements equal to 1. In Supplementary Materials S.5, we derive the covariance decomposition for the hybrid model.

Separability	Model	Covariance Structure
Additive	$Y_i(s,t) = \mu(s,t) + C_i$ $+U_i(s) + V_i(t)$	$\mathbf{\Sigma}_Y = V_C\mathbf{J}_S \otimes \mathbf{J}_T + \mathbf{\Sigma}_U \otimes \mathbf{J}_T$ $+\mathbf{J}_S \otimes \mathbf{\Sigma}_V$
Multiplicative	$Y_i(s,t) = \mu(s,t) + U_i(s)V_i(t)$	$\mathbf{\Sigma}_Y = \mathbf{\Sigma}_U \otimes \mathbf{\Sigma}_V$
Hybrid	$Y_i(s,t) = \mu(s,t) + C_i + U_{1i}(s)$ $+V_{1i}(t) + U_{2i}(s)V_{2i}(t)$	$\mathbf{\Sigma}_Y = V_C\mathbf{J}_S \otimes \mathbf{J}_T + \mathbf{\Sigma}_{U_1} \otimes \mathbf{J}_T$ $+\mathbf{J}_S \otimes \mathbf{\Sigma}_{V_1} + \mathbf{\Sigma}_{U_2} \otimes \mathbf{\Sigma}_{V_2}$

Table 3.1: Three types of separable spatio-temporal models.

3.3 Separable two-way matrix-variate PCA

Our interest centers on obtaining PCA decompositions of the covariance matrices for the latent processes in models (3.1), (3.2) and (3.3). The idea is that PCA could strongly reduce the complexity of modeling by identifying only the main directions of variation in row and column spaces. There are three drawbacks to performing brute force PCA directly on the original matrix $Y_i(s,t)$: 1) it is often difficult to visualize and interpret each principal component as each PC is a two-way matrix; 2) obtaining two-way PCs is computationally intensive due to the large dimensionality of the matrix; and 3) calculating the principal

components will not benefit from the known underlying two-way structure, which could result in unnecessarily complex decompositions.

In the next sections, we introduce PCA methods which account for specific separability assumptions.

3.3.1 Eigenvectors and eigenvalues

In this section we consider estimation of eigenvectors and eigenvalues. For simplicity we discuss only the hybrid model (3.3); analogous, but simpler, approaches can easily be derived for (3.1) and (3.2). Consider model (3.3)

$$Y_i(s,t) = \mu(s,t) + C_i + U_{1i}(s) + V_{1i}(t) + U_{2i}(s)V_{2i}(t).$$

Using Karhunen-Loève transforms (KLT) (Bosq, 2000) for vectors $U_{1i}(s), V_{1i}(t)$ and $U_{2i}(s), V_{2i}(t)$, we obtain

$$U_{1i}(s) = \sum_{k=1}^{S} \xi_{ik}^{U_1} \phi_k^{U_1}(s),\ V_{1i}(t) = \sum_{k=1}^{T} \eta_{ik}^{V_1} \psi_k^{V_1}(t), \tag{3.4}$$

$$U_{2i}(s) = \sum_{k=1}^{S} \xi_{ik}^{U_2} \phi_k^{U_2}(s),\ V_{2i}(t) = \sum_{k=1}^{T} \eta_{ik}^{V_2} \psi_k^{V_2}(t). \tag{3.5}$$

Here the principal scores $\xi_{ik}^{U_j}, \eta_{ik}^{V_j}$ are sequences of real zero-mean random variables such

that

$$E\left[\xi_{ik_1}^{U_j}\xi_{ik_2}^{U_j}\right] = \lambda_{k_1}^{U_j}\delta_{k_1,k_2}; \quad E\left[\eta_{ik_1}^{V_j}\eta_{ik_2}^{V_j}\right] = \lambda_{k_1}^{V_j}\delta_{k_1,k_2}; \quad k_1, k_2 \in \mathbb{N},$$

and $\lambda_k^{U_j}, \lambda_k^{V_j}$ and eigenvectors $\phi_k^{U_j}(s)$, $\psi_k^{V_j}(t)$ are defined such that

$$\sum_{s_1} \mathbf{\Sigma}_{U_j}(s_1, s_2)\phi_k^{U_j}(s_1) = \lambda_k^{U_j}\phi_k^{U_j}(s_2), \quad \sum_{t_1} \mathbf{\Sigma}_{V_j}(t_1, t_2)\psi_k^{V_j}(t_1) = \lambda_k^{V_j}\psi_k^{V_j}(t_2)$$

and

$$\sum_{s} \phi_{k_1}^{U_j}(s)\phi_{k_2}^{U_j}(s) = \delta_{k_1,k_2}, \quad \sum_{s} \psi_{k_1}^{V_j}(t)\psi_{k_2}^{V_j}(t) = \delta_{k_1,k_2}.$$

Plugging the series in the original hybrid model, we obtain

$$Y_i(s,t) = \mu(s,t) + C_i + \sum_{k=1}^{S} \xi_{ik}^{U_1}\phi_k^{U_1}(s) + \sum_{k=1}^{T} \eta_{ik}^{V_1}\psi_k^{V_1}(t) + \sum_{k_1=1}^{S}\sum_{k_2=1}^{T} \xi_{ik_1}^{U_2}\eta_{ik_2}^{V_1}\phi_{k_1}^{U_2}(s)\psi_{k_2}^{V_2}(t).$$

We make the following assumptions:

a. $U_{1i}(s), V_{1i}(t), U_{2i}(s)$ and $V_{2i}(t)$ have mean zero and finite second moments.

b. C_i is uncorrelated with $U_{1i}(s), V_{1i}(t), U_{2i}(s)$ and $V_{2i}(t)$.

c. The row-specific vectors $U_{1i}(s), U_{2i}(s)$ are uncorrelated with the column-specific vectors $V_{1i}(s), V_{2i}(s)$.

Assumption a. is the condition for KLT. Assumptions b. and c. are the identifiability conditions H.3, H.4 and H.5 for the hybird model and provided in Supplementary Materials

S.4. Assumption c. guarantees that the row space (i.e. $U_{1i}(s), U_{2i}(s)$) is separable from the column space (i.e. $V_{1i}(t), V_{2i}(t)$).

3.3.2 Model estimation

We focus first on estimating the eigenvalues and eigenvectors in the hybrid separable model under assumptions a.– c. Let $\Sigma_Y(s_1, s_2, t_1, t_2) = \text{Cov}\{Y_i(s_1, t_1), Y_i(s_2, t_2)\}$ be the overall covariance, and $\Sigma_{U_j}(s_1, s_2) = \text{Cov}\{U_{ji}(s_1), U_{ji}(s_2)\}$ and $\Sigma_{V_j}(t_1, t_2) = \text{Cov}\{V_{ji}(t_1), V_{ji}(t_2)\}$ be the covariance matrices for the row-specific and column-specific processes, respectively. From Table 3.1 and Equation (3.4), if follows that

$$\begin{aligned}\Sigma_Y(s_1, s_2, t_1, t_2) &= V_C + \Sigma_{U_1}(s_1, s_2) + \Sigma_{V_1}(t_1, t_2) + \Sigma_{U_2}(s_1, s_2)\Sigma_{V_2}(t_1, t_2) \\ &= V_C + \sum_{k=1}^{S} \lambda_k^{U_1} \phi_k^{U_1}(s_1)\phi_k^{U_1}(s_2) + \sum_{k=1}^{T} \lambda_k^{V_1} \psi_k^{V_1}(t_1)\psi_k^{V_1}(t_2) \\ &\quad + \sum_{k_1=1}^{S}\sum_{k_2=1}^{T} \lambda_{k_1}^{U_2}\lambda_{k_2}^{V_1} \phi_{k_1}^{U_2}(s_1)\phi_{k_1}^{U_2}(s_2)\psi_{k_2}^{V_2}(t_1)\psi_{k_2}^{V_2}(t_2)\end{aligned}$$

Our general procedure consists of following 4 steps.

Step 1. estimate the mean and covariance $\hat{\mu}(s,t)$, $\widehat{\Sigma}_{U_j}(s_1, s_2)$ and $\widehat{\Sigma}_{V_j}(t_1, t_2)$ using the method of moments.

Step 2. perform SVD on $\widehat{\Sigma}_{U_j}(s_1, s_2)$ to obtain estimates of eigenvalues $\widehat{\lambda}_k^{U_j}$ and eigenvectors $\widehat{\phi}_k^{U_j}(s)$.

Step 3. perform SVD on $\widehat{\Sigma}_{V_j}(t_1, t_2)$ to obtain estimates of eigenvalues $\widehat{\lambda}_k^{V_j}$ and eigenvectors $\widehat{\psi}_k^{V_j}(t)$.

Step 4. estimate the principal scores.

In Step 1, we start by subtracting an estimator of the mean $\mu(s,t)$. In this paper, we use the empirical average. By essentially the same arguments invoked in the classical case of an unstructured covariance matrix (see e.g. Anderson (2003), Section 3.2), it can be shown that for any fixed positive definite $V_C\mathbf{J}_S \otimes \mathbf{J}_T + \mathbf{\Sigma}_{U_1} \otimes \mathbf{J}_T + \mathbf{J}_S \otimes \mathbf{\Sigma}_{V_1} + \mathbf{\Sigma}_{U_2} \otimes \mathbf{\Sigma}_{V_2}$, the empirical mean is the maximum likelihood estimate if normality is assumed. For simplicity, we still denote the demeaned process $\{Y_i(s,t) - \hat{\mu}(s,t)\}$ as $Y_i(s,t)$. Let $\mathbf{Y}_i$ be the $S \times T$ matrix, i.e.

$$\mathbf{Y}_i = \begin{pmatrix} Y_i(s_1,t_1) & Y_i(s_1,t_2) & ... & Y_i(s_1,t_T) \\ Y_i(s_2,t_1) & Y_i(s_2,t_2) & ... & Y_i(s_2,t_T) \\ ... & ... & ... & ... \\ Y_i(s_S,t_1) & Y_i(s_S,t_2) & ... & Y_i(s_S,t_T) \end{pmatrix}$$

where $i = 1, 2, ..., N$. Similarly, we define $\mathbf{U_{1}}_i, \mathbf{U_{2}}_i, \mathbf{V_{1}}_i$ and $\mathbf{V_{2}}_i$ to be the vectors of $U_{1i}(s), U_{2i}(s), V_{1i}(t)$ and $V_{2i}(t)$. Therefore,

$$\mathbf{Y}_i = C_i \mathbf{1_S} \mathbf{1_T}^T + \mathbf{U_{1}}_i \mathbf{1_T}^T + \mathbf{1_S} \mathbf{V_{1}}_i^T + \mathbf{U_{2}}_i \mathbf{V_{2}}_i^T.$$

The following equalities hold:

$$
\begin{aligned}
E[\mathbf{1_S}^T\mathbf{Y}_i\mathbf{1_T}\mathbf{1_T}^T\mathbf{Y}_i^T\mathbf{1_S}] &= S^2T^2V_C \\
E[\mathbf{Y}_i\mathbf{1_T}\mathbf{1_T}^T\mathbf{Y}^T] &= T^2V_C\mathbf{1_S}\mathbf{1_S}^T + T^2\mathbf{\Sigma_{U_1}} \\
E[\mathbf{Y}_i^T\mathbf{1_S}\mathbf{1_S}^T\mathbf{Y}] &= S^2V_C\mathbf{1_T}\mathbf{1_T}^T + S^2\mathbf{\Sigma_{V_1}} \\
E[\mathbf{Y}_i\mathbf{Y}_i^T] &= TV_C\mathbf{1_S}\mathbf{1_S}^T + T\mathbf{\Sigma_{U_1}} + \text{trace}(\mathbf{\Sigma_{V_1}})\mathbf{1_S}\mathbf{1_S}^T + \text{trace}(\mathbf{\Sigma_{V_2}})\mathbf{\Sigma_{U_2}} \\
E[\mathbf{Y}_i^T\mathbf{Y}_i] &= SV_C\mathbf{1_S}\mathbf{1_S}^T + \text{trace}(\mathbf{\Sigma_{U_1}})\mathbf{1_T}\mathbf{1_T}^T + S\mathbf{\Sigma_{V_1}} + \text{trace}(\mathbf{\Sigma_{U_2}})\mathbf{\Sigma_{V_2}}
\end{aligned}
$$

and can be used to obtain the following MoM estimates for the covariance functions

$$
\begin{aligned}
\widehat{V}_C &= \frac{1}{NS^2T^2}\sum_{i=1}^{N}\mathbf{1_S}^T\mathbf{Y}_i\mathbf{1_T}\mathbf{1_T}^T\mathbf{Y}_i^T\mathbf{1_S} \\
\widehat{\mathbf{\Sigma}}_{\mathbf{U_1}} &= \frac{1}{T^2}\left\{\frac{1}{N}\sum_{i=1}^{N}\mathbf{Y}_i\mathbf{1_T}\mathbf{1_T}^T\mathbf{Y}^T - T^2\widehat{V}_C\mathbf{1_S}\mathbf{1_S}^T\right\} \\
\widehat{\mathbf{\Sigma}}_{\mathbf{V_1}} &= \frac{1}{S^2}\left\{\frac{1}{N}\sum_{i=1}^{N}\mathbf{Y}_i^T\mathbf{1_S}\mathbf{1_S}^T\mathbf{Y} - S^2\widehat{V}_C\mathbf{1_T}\mathbf{1_T}^T\right\} \\
\widehat{\mathbf{\Sigma}}_{\mathbf{U_2}} &= \frac{1}{\widehat{D}}\left\{\frac{1}{N}\sum_{i=1}^{N}\mathbf{Y}_i\mathbf{Y}_i^T - T\widehat{V}_C\mathbf{1_S}\mathbf{1_S}^T - T\widehat{\mathbf{\Sigma}}_{\mathbf{U_1}} - \text{trace}(\widehat{\mathbf{\Sigma}}_{\mathbf{V_1}})\mathbf{1_S}\mathbf{1_S}^T\right\} \\
\widehat{\mathbf{\Sigma}}_{\mathbf{V_2}} &= \frac{1}{\text{trace}(\widehat{\mathbf{\Sigma}}_{\mathbf{U_2}})}\left\{\frac{1}{N}\sum_{i=1}^{N}\mathbf{Y}_i^T\mathbf{Y}_i - S\widehat{V}_C\mathbf{1_T}\mathbf{1_T}^T - \text{trace}(\widehat{\mathbf{\Sigma}}_{\mathbf{U_1}})\mathbf{1_T}\mathbf{1_T}^T - S\widehat{\mathbf{\Sigma}}_{\mathbf{V_1}}\right\}
\end{aligned}
$$

where $\hat{D}$ is the normalizing constant to ensure identifiability condition H.8.

In the traditional PCA literature (Jolliffe, 2002), estimation often involves calculating the empirical estimate of the overall covariance function $\Sigma_Y(s_1, s_2, t_1, t_2)$, which requires storing of $O(S^2 \times T^2)$ parameters. When S or T is large, this population level estimate is

not computationally feasible. Instead, by taking advantage of separability, one requires the storage of only $O(S^2 + T^2)$ parameters.

Once Step 1 is implemented, Steps 2 and 3 are routine. In Step 4, we are using projection methods to estimate the scores. It remains unclear how to estimate the individual scores $\xi_{ik}^{U_2}$ and $\eta_{ik}^{V_2}$. In this paper we focus on estimating the products $\xi_{ik_1}^{U_2}\eta_{ik_2}^{V_2}$, which is a much simpler problem. All computational details can be found in Supplementary Materials S.6.

Choosing the number of principal components is an important practical problem without a theoretically satisfactory solution. Two practical alternatives are to use cross validation (Rice and Silverman, 1991) or Akaike Information Criterion (Akaike, 1974; Yao et al., 2005). One might choose an even simpler method for estimating the number of components based on the estimated explained variance; this approach has been used extensively in practice and seems to be the most prevalent approach. More precisely, let P be a threshold and define

$$N_{U_1} = \min\{k : \rho_k^{U_1} \geq P\}$$

where $\rho_k^{U_1} = (\lambda_1^{U_1} + ... + \lambda_k^{U_1})/(\lambda_1^{U_1} + ... + \lambda_S^{U_1})$. In our analysis we used $P = 0.8$. We used a similar method for choosing the number of components for $U_2(s), V_1(t), V_2(t)$. These choices were slightly conservative, but worked well in our simulations and application. However, the threshold should be carefully tuned in particular applications.

A novel aspect of our approach is that it allows us to assess the relative variability ex-

plained by the row-specific process versus the column-specific process. For example, in the additive model, the row-space variability can be quantified by $\sum_{k=1}^{S} \lambda_k^U = \sum_s \text{var}\{U_i(s)\}$ while the column-space variability can be quantified by $\sum_{k=1}^{T} \lambda_k^V = \sum_t \text{var}\{V_i(t)\}$. A natural measure of the proportion variability explained by the column-specific process is

$$r_T = \frac{T\sum_{k=1}^{S}\lambda_k^U}{T\sum_{k=1}^{S}\lambda_k^U + S\sum_{k=1}^{T}\lambda_k^V} = \frac{\sum_t\sum_s \text{var}\{U_i(s)\}}{\sum_t\sum_s \text{var}\{U_i(s)\} + \sum_s\sum_t \text{var}\{V_i(t)\}}$$

Similarly, we can define the proportion of variability explained by the additive component versus the multiplicative component in the hybrid model as

$$\begin{aligned} r_M &= \frac{\sum_{k=1}^{S}\lambda_k^{U_2}\sum_{k=1}^{T}\lambda_k^{V_2}}{T\sum_{k=1}^{S}\lambda_k^{U_1} + S\sum_{k=1}^{T}\lambda_k^{V_1} + \sum_{k=1}^{S}\lambda_k^{U_2}\sum_{k=1}^{T}\lambda_k^{V_2}} \\ &= \frac{\sum_s\sum_t \text{var}\{U_{2i}(s)\}\text{var}\{V_{2i}(t)\}}{\sum_s\sum_t \text{var}\{U_{2i}(s)\} + \sum_s\sum_t \text{var}\{V_{2i}(t)\} + \sum_s\sum_t \text{var}\{U_{2i}(s)\}\text{var}\{V_{2i}(t)\}} \end{aligned}$$

3.4 Two-way matrix data with white noise

So far we have assumed that data are measured without noise or that the noise can be absorbed into one of the latent processes. However, the methods can easily be extended to data contaminated by white noise $\varepsilon_i(s,t) \sim N(0, \sigma_\varepsilon^2)$ by defining a the new symmetric covariance matrix $\tilde{\Sigma}_Y$ such that $\tilde{\Sigma}_Y(s_1, s_2, t_1, t_2) = \Sigma_Y(s_1, s_2, t_1, t_2) + \sigma_\varepsilon^2\delta_{s_1s_2}\delta_{t_1t_2}$.

We start with smoothness assumptions to ensure the identifiability of the noisy version of models (3.1), (3.2) and (3.3), i.e. the covariance matrix of either row-specific processes

or column-specific processes are smooth bivariate functions. Table 3.2 shows the specific additional assumption for each theorem. In our study, the assumption is plausible because the signals are recorded in a continuous time domain, which induces smoothness in the column space.

Theorem	Extra assumption	Assumption statement
Theorem 1	A.5	$\mathrm{Cov}\{U_i(s_1), U_i(s_2)\}$ or $\mathrm{Cov}\{V_i(t_1), V_i(t_2)\}$ is a smooth bivariate function.
Theorem 2	M.6	$\mathrm{Cov}\{V_i(t_1), V_i(t_2)\}$ is a smooth bivariate function.
Theorem 3	H.10	$\mathrm{Cov}\{U_{1i}(s_1), U_{1i}(s_2)\}$ or $\mathrm{Cov}\{V_{1i}(t_1), V_{1i}(t_2)\}$ is a smooth bivariate function.

Table 3.2: Additional assumption for models with white noise to ensure identifiability of models (3.1), (3.2) and (3.3), respectively.

We can then apply off-diagonal smoothing (Staniswalis and Lee, 1998; Greven et al., 2010) on $\mathbf{\Sigma_V}$. For example, in model (3.3) with white noise $\varepsilon_i(s,t)$ we have

$$E[\mathbf{Y}_i^T \mathbf{1_S} \mathbf{1_S}^T \mathbf{Y}_i] = S^2 V_C \mathbf{1_T} \mathbf{1_T}^T + S^2 \mathbf{\Sigma_{V_1}} + S\sigma_\varepsilon^2 \mathbf{I_T}.$$

By off-diagonal smoothing on $\frac{1}{N}\sum_{i=1}^N \mathbf{Y}_i^T \mathbf{1_S} \mathbf{1_S}^T \mathbf{Y}_i$, we obtain an estimate of σ_ε^2.

In practice the off-diagonal smoothing method sometimes overestimates the noise variance, especially when the signal-to-noise ratio is low. This may cause the estimated covariance matrix to have negative eigenvalues. One alternative is to set the upper bound for the estimated noise variance by identifying the smallest eigenvalue of the original covariance matrix. More precisely, the eigenvalues of the covariance matrix $\tilde{\mathbf{\Sigma}}_\mathbf{U}$ or $\tilde{\mathbf{\Sigma}}_\mathbf{V}$ with noise can be decomposed as the eigenvalues of the noise free $\mathbf{\Sigma_U}$ or $\mathbf{\Sigma_V}$ and the constant variance of

the white noise. By looking at the smallest eigenvalue of the estimated covariance matrix (where the non-noise portion is small), we can provide the upper bound for the variance of the white noise. To the best of our knowledge, this is a new technique, which stabilizes the noise error variance estimators and improves on the original method of Staniswalis and Lee (1998).

Similar estimation techniques can be applied to models (3.1) and (3.2). In practice, if σ_ε^2/S or σ_ε^2/T is small, relative to $V_C + \Sigma_V(t,t')$ or $V_C + \Sigma_U(s,s')$, we may simply ignore the noise term (that is, setting $\sigma_\varepsilon^2 = 0$). This is because 1) the off-diagonal smoothing is not sensitive enough to separate the noise term; 2) the bias for $\Sigma_U(s,s')$ or $\Sigma_V(t,t')$ is at most σ_ε^2/S or σ_ε^2/T, which are typically negligible due to the large S or T in high-dimensional cases.

3.5 Simulation

To better understand the performance of the proposed two-way PCA in practice, we conduct simulation studies for the additive model (3.1), multiplicative model (3.2) and hybrid model (3.3).

We generated data from the model

$$\begin{cases} Y_i(s,t) = C_i + \sum_{k_1=1}^{K_{U_1}} \xi_{ik_1}^{U_1}\phi_{k_1}^{U_1}(s) + \sum_{k_1'=1}^{K_{V_1}} \eta_{ik_1'}^{V_1}\psi_{k_1'}^{V_1}(t) \\ \qquad\qquad + \left[\sum_{k_2=1}^{K_{U_2}} \xi_{ik_2}^{U_2}\phi_{k_2}^{U_2}(s)\right]\left[\sum_{k_2'=1}^{K_{V_2}} \eta_{ik_2'}\psi_{k_2'}^{V_2}(t)\right] + \varepsilon_i(s,t),\ s\in\mathcal{S}, t\in\mathcal{T} \\ C_i \sim N(0,\sigma_C^2), \xi_{ik_1} \sim N(0,\lambda_{k_1}^{U_1}), \xi_{ik_2} \sim N(0,\lambda_{k_2}^{U_2}) \\ \eta_{ik_1'} \sim N(0,\lambda_{k_1'}^{V_1}), \eta_{ik_2'} \sim N(0,\lambda_{k_2'}^{V_2}), \varepsilon_i(s,t) \sim N(0,\sigma_\varepsilon^2) \end{cases}$$

where $\phi_{k_1}^{U_1}(s), \psi_{k_1'}^{V_1}(t), \phi_{k_2}^{U_2}(s)$ and $\psi_{k_2'}^{V_2}(t)$ are the eigenvectors, $\xi_{ik_1}^{U_1}, \xi_{ik_2}^{U_2}, \eta_{ik_1'}^{V_1}$ and $\eta_{ik_2'}^{V_2}$ are the PC scores, $K_{U_1}, K_{V_1}, K_{U_2}$ and K_{V_2} are the number of components, grids $\mathcal{S} = \{1/30, 2/30,$ ..., $1\}$, $\mathcal{T} = \{1/20, 2/20, ..., 1\}$ and N is the sample size.

We set $K_{U_1} = 4$, $K_{V_1} = 3$, $K_{U_2} = K_{V_2} = 0$ for the additive model, $K_{U_1} = K_{V_1} = 0$, $K_{U_2} = 4, K_{V_2} = 3$ for the multiplicative model and $K_{U_1} = 4$, $K_{V_1} = 3$, $K_{U_2} = K_{V_2} = 2$ for the hybrid model. We simulate random scores $C_i, \xi_{ik_1}^{U_1}, \xi_{ik_2}^{U_2}, \eta_{ik_1'}^{V_1}$ and $\eta_{ik_2'}^{V_2}$ and white noise $\varepsilon_i(s,t)$ independently, where σ_C^2 is fixed at 1. The values of λ_k^W are fixed at $\frac{1}{2}^{k-1}$, while σ_ε is selected to match different levels of the signal-to-noise. Our simulation experiment varies two parameters: the sample size, N, and the signal-to-noise ratio, $\sigma_y^2/\sigma_\varepsilon^2$. Supplementary Materials S.7 provides details of the design, including the definitions of the eigenvectors. For each parameter setting, we performed 100 replications. Figure 3.3 compares the estimated PCs for the hybrid model with the true eigenvectors, and shows the accurate performance of our method. Estimation accuracy can be quantified by the mean square error MSE $= \frac{1}{n_x}\sum_{i=1}^{n_x}\left\{f(x_i) - \hat{f}(x_i)\right\}^2$, where $f(x)$ and $\hat{f}(x)$ are the true and estimated principal components and n_x is the size of the grid of x_i's. MSEs for hybrid model

are summarized in Table 3.3. Detailed simulation results for additive and multiplicative

$\sigma_\varepsilon^2/\sigma_y^2$	0	0.1	1	10	100
$\phi_1^{U_1}$	0.0070	0.0065	0.0070	0.0143	0.2210
$\phi_2^{U_1}$	0.0134	0.0115	0.0130	0.0327	0.7761
$\phi_3^{U_1}$	0.0169	0.0187	0.0222	0.0704	1.7663
$\phi_4^{U_1}$	0.0182	0.0187	0.0254	0.1735	1.8831
$\psi_1^{V_1}$	0.0071	0.0067	0.0074	0.0141	0.1335
$\psi_2^{V_1}$	0.0124	0.0122	0.0132	0.0268	0.5338
$\psi_3^{V_1}$	0.0125	0.0131	0.0146	0.0405	1.6790
$\phi_1^{U_2}$	0.0096	0.0104	0.0113	0.0183	0.8929
$\phi_2^{U_2}$	0.0105	0.0110	0.0124	0.0266	1.6112
$\psi_1^{V_2}$	0.0104	0.0104	0.0110	0.0162	0.9476
$\psi_2^{V_2}$	0.0131	0.0122	0.0130	0.0227	1.5170

Table 3.3: Average MSE of the principal components under different signal-to-noise ratio for hybrid model.

models can be found in the supplementary materials.

3.6 Data application

We apply our proposed method to data from the fMRI study of thermal pain described in the Introduction. For simplicity we discuss only the hybrid model results.

Our analysis included 20 subjects measured over 21 brain regions with 45–52 trials per subject, each consisting of 23 time points. Figure 3.4 displays the estimated overall mean fMRI signal, and the mean signal over time and space. In the left panel, the overall mean fMRI signal is calculated as the empirical average across all subjects and all trials. The middle panel shows the mean temporal signal, which is the simple average in the temporal domain across all the brain regions and all trials. The upward trend in the temporal pattern

is likely associated with the cumulative effect of the thermal pain applied at the beginning of each trial. The right panel displays the spatial mean, which is the average across all time points and trials. Compared to the temporal mean, the variation in the spatial mean is relatively small.

3.6.1 Variation analysis and principal component analysis

For the temporal (column-specific) term $V_{1i}(t)$ in the additive component $U_{1i}(s) + V_{1i}(t)$, Table 3.4 displays the estimated eigenvalues, indicating that most of the temporal information is contained in the first 3–4 components. The first three eigenvalues explain 35%, 24% and 11% of the variation, respectively—together, over 71% of the temporal variability in the additive component. The upper-right panel of Figure 3.5 displays the first four eigenvectors. The first principal component (black line) corresponds to activation related to the subjects' initial reaction to the thermal stimulus. The second principal component (red line) peaks between 20 and 24 seconds after the start of the trial. This corresponds to the first 4 seconds following the end of heat application and represents the response to the thermal stimulation. The delayed effect is due to the delayed nature of brain hemodynamics, which peaks roughly 6 seconds after peak neuronal activation, and is consistent with timings of other fMRI experiments (Lindquist, 2008). The fourth principal component (blue line) has two peaks: one during the 14–18 second interval when the stimulus ended, and the other during the 38–44 second interval, which immediately precedes reporting. Thus, this

Spatial eigenvalues in the additive part						
Component	1	2	3			
eigenvalue	0.45	0.075	0.011			
percent var	83.83	13.98	2.18			
cum percent var	83.83	97.81	99.98			
Temporal eigenvalues in the additive part						
Component	1	2	3	4	5	6
eigenvalue	3.72	2.58	1.19	1.00	0.74	0.54
percent var	35.42	24.54	11.32	9.54	7.04	5.13
cum percent var	35.42	59.96	71.28	80.83	87.87	93.00
Spatial eigenvalues in the multiplicative part						
Component	1	2	3	4	5	
eigenvalue	0.78	0.23	0.17	0.13	0.11	
percent var	49.48	14.34	10.93	8.23	7.15	
cum percent var	49.48	63.82	74.76	83.00	90.14	
Temporal eigenvalues in the multiplicative part						
Component	1	2	3	4		
eigenvalue	0.24	0.20	0.08	0.06		
percent var	37.74	30.91	13.08	10.31		
cum percent var	37.74	68.66	81.74	92.06		

Table 3.4: Estimated eigenvalues on temporal, spatial and multiplicative term. "percent var" stands for the percentage of variance explained by the component, and "cum percent var" means the cumulative percentage of variance explained

component may be linked to the initial pain response (first peak) and pain recall (second peak).

The spatial (row-specific) term $U_{1i}(s)$ in the additive component has even fewer directions of variation. Indeed, 99% of the variability is explained by the first three principal components. The concentration of variability most likely reflects the larger spatial homogeneity among the "selected" brain regions. The upper-left panel of Figure 3.5 displays the first three spatial components, where the x-axis indexes the brain locations. Notation of regions in the Figure are 1 through 21, while Table S.1 in Supplementary Materials S.1 pro-

vides the name of these regions. The first principal component loads negatively on regions in the anterior insula (AINS), the dorsal anterior cingulate cortex (dACC) and the corpus callosum. These are all regions that have been shown to mediate the relationship between thermal stimuli and pain response consistently across subjects (Atlas et al., 2014). The second principal component loads positively on the insula, thalamus and dACC, and negatively on the parahippocampal gyrus (PHG) and inferior frontal gyrus (IFG). Interestingly, in previous analyses activation in the former regions all showed a positive linear effect of applied temperature, while the PHG showed a negative effect (Atlas et al., 2014). Finally, the third principal component loads positively on IFG, occipital gyrus and corpus callosum and negatively on the second somatosensory area (SII).

In the lower two panels of Figure 3.5, we display the row-specific and column-specific principal components for the multiplicative component $U_{2i}(s)V_{2i}(t)$. The leading principal components are very similar to those in the additive component.

The proportion of variability explained by the column-specific term in the additive component, r_T, was defined in Section 3.3.1. In the thermal pain study we estimate $\hat{r}_T = 0.594$, that is, 59% of variability in the thermal pain data is attributable to the temporal domain in the additive component. We can also calculate the proportion of variability explained by the multiplicative component, $\hat{r}_M = 0.373$. It indicates a portion of 37% of variability comes from the higher order interaction.

One may be interested in formulating a statistical test on the significance of each term in the hybrid model. For example, we can test the null hypothesis $H_0 : r_M = 0$ ver-

sus $H_A : r_M > 0$. We used a parametric bootstrap, where we fitted the hybrid model to the thermal pain data under H_0 and kept the first three row-specific principal components and the first six column-specific components in the additive component. Based on the estimated model, we generated bootstrap samples and extracted eigenvalues and eigenfunctions. Based on 1000 bootstrap samples, the 95% confidence interval for r_M under null hypothesis is $(0.0103, 0.0115)$, which does not contain $\hat{r}_M = 0.373$. Thus, we can reject the null hypothesis that the multiplicative component in the hybrid model is not statistically significant. We also carried out a nonparametric bootstrap by resampling the subjects, resulting in a 95% confidence interval for $\hat{r}_M$ is $(0.302, 0.469)$, which does not include 0.

3.6.2 Distribution of PC scores

One of the main goals of PCA is dimensionality reduction. For example, the high dimensional column-specific (temporal) process in fMRI has a representation in terms of 4-dimensional vectors of scores. This low-dimensional representation can then be used in subsequent analyses, by using scores either as covariates or outcomes.

As discussed in Section 3.3.1, we estimated the principal component scores using the projection method (Di et al., 2009). Figure 3.6 displays the distribution of the estimated scores for $U_{1i}(s)$, $V_{1i}(t)$ and $U_{2i}(s)V_{2i}(t)$. The upper left panel shows the scatterplot of the first and second PC scores for $V_{1i}(t)$. The first component explains 43.32% of the variation and corresponds to the hemodynamic response to the initial portion of the pain stimuli.

Note that the first component scores have a much wider range than the second component scores, which is consistent with the much larger variability explained by the first component. The other panel shows the first few PC scores for $U_{1i}(s)$, $V_{1i}(t)$ and $U_{2i}(s)V_{2i}(t)$ versus the stimulation setting covariate. Figure 3.6 indicates that the hot temperature settings tend to have a higher second PC scores for $V_{1i}(t)$ (with p value less than 0.05 under a two-sample t test).

3.6.3 Association between component scores and pain rating

In this section we analyze the association between the fMRI signal and pain rating. Consider the linear mixed effect regression model

$$Y_{ij} = \beta_0 + \sum_s \sum_t \beta(s,t) X_{ij}(s,t) + Z_{ij}\gamma + b_{0i} + \sum_s \sum_t b_i(s,t) X_{ij}(s,t) + Z_{ij} a_i + \varepsilon_{ij} \quad (3.6)$$

where Y_{ij} is the log pain rating for subject i at trial j, Z_{ij} is an indicator of applied temperature, and $X_{ij}(s,t)$ is the fMRI signal for subject i at trial j. Since we can approximate

$X_{ij}(s,t)$ by KLT, the equation (3.6) becomes

$$Y_{ij} = \beta_0 + C_{ij}\beta^C + \sum_{k_1=1}^{K_{U_1}} \beta_{k_1}^{\xi_{U_1}} \xi_{ijk_1}^{U_1} + \sum_{k_1'=1}^{K_{V_1}} \beta_{k_1'}^{\eta_{V_1}} \eta_{ijk_1'}^{V_1} + \sum_{k_2=1}^{K_{U_2}} \sum_{k_2'=1}^{K_{V_2}} \beta_{k_2k_2'}^{M} \xi_{ijk_2}^{U_2} \eta_{ijk_2'}^{V_2}$$
$$+ b_{0i} + C_{ij} b_i^C + \sum_{k_1=1}^{K_{U_1}} b_{ik_1}^{\xi_{U_1}} \xi_{ijk_1}^{U_1} + \sum_{k_1'=1}^{K_{V_1}} b_{ik_1'}^{\eta_{V_1}} \eta_{ijk_1'}^{V_1} + \sum_{k_2=1}^{K_{U_2}} \sum_{k_2'=1}^{K_{V_2}} b_{ik_2k_2'}^{M} \xi_{ijk_2}^{U_2} \eta_{ijk_2'}^{V_2} + \varepsilon_{ij}$$

where $\beta^C = \sum_s \sum_t \beta(s,t)$, $\beta_{k_1}^{\xi_{U_1}} = \sum_s \sum_t \beta(s,t)\phi_{k_1}^{U_1}(s)$, $\beta_{k_1'}^{\eta_{V_1}} = \sum_s \sum_t \beta(s,t)\psi_{k_1'}^{V_1}(t)$, $\beta_{k_2k_2'}^{M} = \sum_s \sum_t \beta(s,t)\phi_{k_2}^{U_2}(s)\phi_{k_2'}^{V_2}(t)$, $b_i^C = \sum_s \sum_t b_i(s,t)$, $b_{ik_1}^{\xi_{U_1}} = \sum_s \sum_t b_i(s,t)\phi_{k_1}^{U_1}(s)$, $b_{ik_1'}^{\eta_{V_1}} = \sum_s \sum_t b_i(s,t)\psi_{k_1'}^{V_1}(t)$ and $b_{ik_2k_2'}^{M} = \sum_s \sum_t b_i(s,t)\phi_{k_2}^{U_2}(s)\phi_{k_2'}^{V_2}(t)$. This is a linear mixed effect model with PC scores as the covariates.

We fit several models with different random effects, with results summarized in Table 3.5. Model 1 is a linear model with no random effect. Model 2 is a mixed effect model with only random intercept. Model 3 has both random intercept and random effect of C_{ij} scores. All three models indicate that 1) temperature has a strong positive association with pain rating; 2) the second and fourth principal component scores are positively associated with the pain rating.

The positive correlation between temperature and pain rating is quite intuitive. In the high heat setting (i.e. TempSetting=1), subjects will tend to report higher pain scores. The second principal component represents the hemodynamic response to the entire thermal stimulus. The positive coefficient indicates that people experiencing high-rating pain tend to have a higher fMRI intensity in response to the stimulus. The fourth principal component

	Model 1		Model 2		Model 3	
	coefficient	p-value	coefficient	p-value	coefficient	p-value
TempSetting	0.5860	$<$ 2e-16	0.6226	$<$ 2e-16	0.6182	$<$ 2e-16
constant	-0.0440	0.5328	-0.0660	0.6895	-0.0641	0.3695
spatial 1	-0.0114	0.0467	0.0066	0.0325	-0.0084	0.0563
temporal 1	0.0007	0.0082	0.0003	0.0058	0.0003	0.0126
temporal 2	0.0014	0.0001	0.0012	0.0001	0.0012	0.0006
temporal 3	-0.0002	0.5207	-0.0003	0.6982	-0.0002	0.8965
temporal 4	0.0019	0.0003	0.0015	0.0012	0.0014	0.0052
multiplicative	-0.0003	0.1031	-0.0001	0.5038	-0.0002	0.3698

Table 3.5: Coefficients for fixed effects for three mixed effect models which regress PCA scores on the pain rating scores.

may represent the pain recall process. Increased activation at this time tends to correspond to higher pain report. Interestingly, the sign of the spatial scores correspond to those seen in neurological signatures previously used to predict physical pain from brain activation (Wager et al., 2013).

3.7 Discussion

This article introduces three types of separable two-way matrix-variate models, using explicit latent process modeling. Identifiability conditions are introduced and method-of-moments estimators are provided for the covariance matrices of all latent processes. Principal component analysis is then used for dimensionality reduction at the level of individual spatial and temporal processes. Methods are applied to data observed with or without white noise. When we applied the method to data from the fMRI study, we distinguished various patterns inherent in the data and quantified the amount of variability captured by the

various components.

The models we proposed here are mainly for exploratory purposes. In other words, goodness-of-fit of the models are not the main focus here. However, we can still use a bootstrap test to formally test the separability assumption on the covariances matrix. We can use the difference of log determinants of the covariance matrix as the test statistics (Lu and Zimmerman, 2005) to test separable covariance matrix under null hypothesis against more general covariance structure under alternative. From the test, unfortunately, our separability models are not statistically significant. However, given that our method was able to extract meaningful and interpretable results and our focus was on exploratory data analysis, we shall not underestimate the usefulness of the proposed method. To improve the robustness of the model, we can extend hybrid model of separability to a model of the type $Y_i(s,t) = h\{U_i(s), V_i(t)\} + \varepsilon_i(s,t)$, where the processes $U_i(s)$ and $V_i(t)$ are uncorrelated and $h(u,v)$ is a specified function. For example the hybrid model could be obtained with $U_i(s) = [U_{1i}(s), U_{2i}(s)]^T$, $V_i(t) = [V_{1i}(t), V_{2i}(t)]^T$ and $h(u,v) = u^T[1,0]^T + v^T[1,0]^T + u^T[0,1]^T v^T[0,1]^T$. Whether or not more complex $h(u,v)$ functions will be useful remains an open problem, though we find the explicit definition of separability to be quite useful.

Besides above generalization of the models, our approaches also suggest other several future directions of research. First, the estimated covariance matrix cannot be guaranteed to be positive definite when the number of subjects is less than the maximum of the dimensions of space and time. One potential solution is to work on some sub-model of the pro-

posed separable models. For example, Fosdick and Hoff (2014) decomposed the covariance matrix to the sum of a reduced-rank matrix and a diagonal matrix and then approximated the reduced-rank matrix by a factor model using MLE. Second, the noise-free version of our method of moments estimators are similar to the two-directional two-dimensional PCA (Zhang and Zhou, 2005), which may suffer from noise contamination. We could address this problem by the off-diagonal smoothing technique mentioned in Section 3.4. Another alternative is to consider a multilinear estimator as described in Hung et al. (2012) by iterative alternating least squares estimation. Third, we did not take the multi-level data structure into account. To solve this problem, we can implement the decomposition technique proposed by Shou et al. (2014) before the spatio-temporal variations are separated. Fourth, we could develop rigorous treatment of noise (Di et al., 2009) as well as address possible sparsity in the functional observations (Di et al., 2014b). Last, a hypothesis testing framework for separable models should be developed. A likelihood ratio test can be proposed if we impose more assumptions in the models (Lu and Zimmerman, 2005). However, even if models do not hold, they can still be useful as an approximation for exploratory purposes.

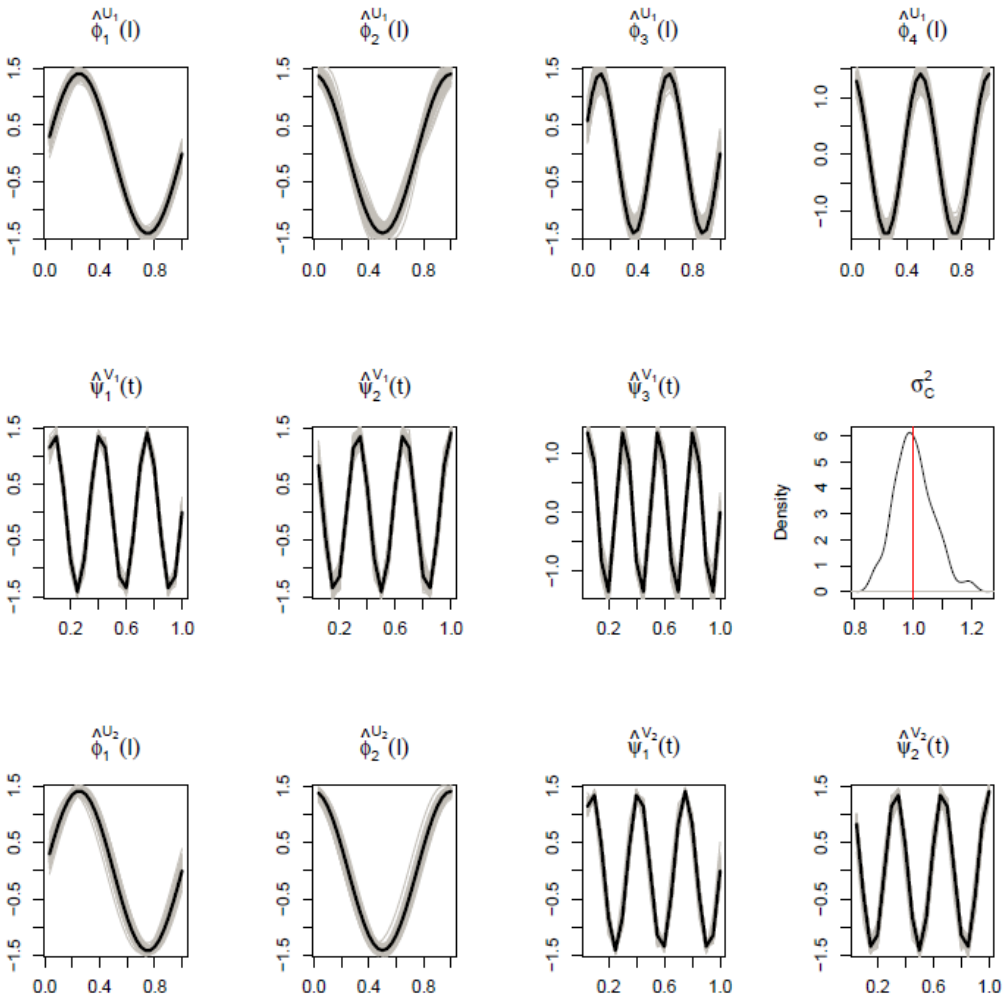

Figure 3.3: The estimated principal components when N=500 in 100 simulations for hybrid model are shown in gray bands. Black curves are the true eigenvectors. The figure also contains the density function of the estimated σ_C^2, plotted with the red vertical line marking the true value.

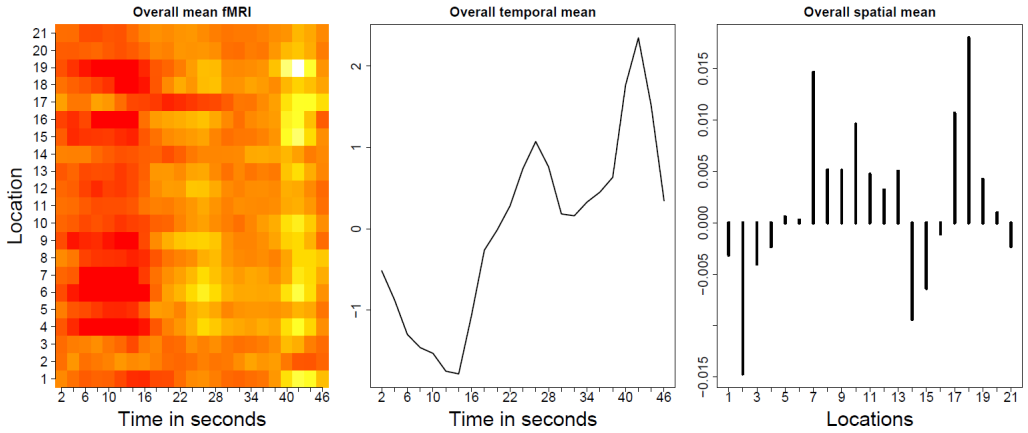

Figure 3.4: Estimated overall, temporal and spatial mean functions for thermal pain fMRI data. The left panel shows the overall mean of fMRI across all subjects and all trials. The x-axis indexes the time course and the y-axis indexes the brain regions (see Table S.1 for more information). The middle panel shows the marginal mean of the temporal signal. The x-axis denotes the time course. The right panel shows the marginal mean of the spatial signal.

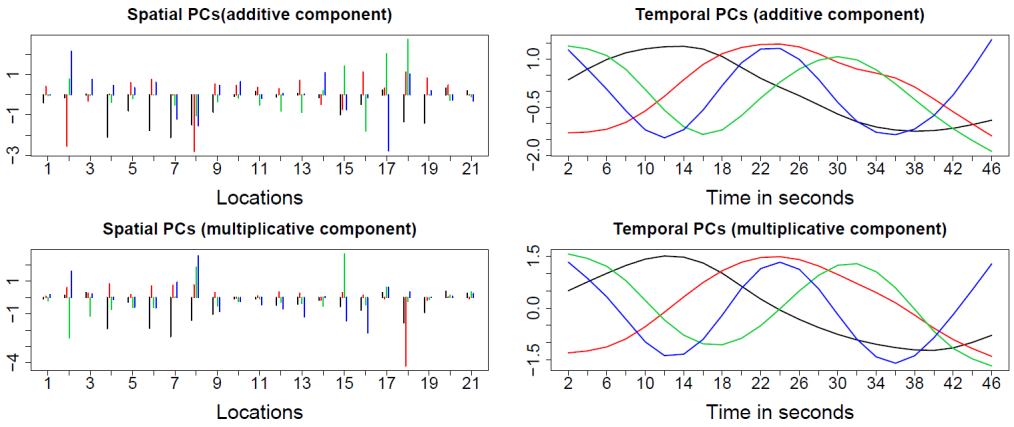

Figure 3.5: Principal components under hybrid model. The left panel shows the row-specific or spatial PCs. The right panel shows the column-specific or temporal PCs. The black, red, green and blue lines stand for the first, second, third and fourth components, respectively. For more information about locations see Table S.1.

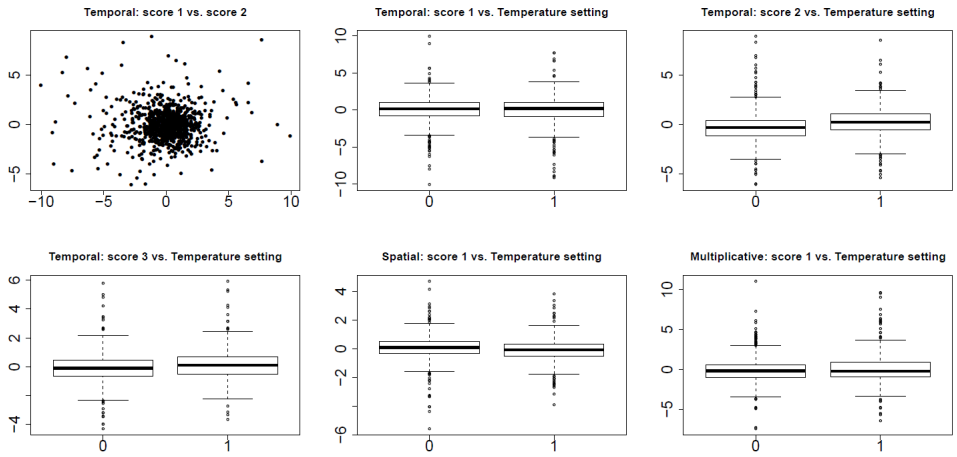

Figure 3.6: Estimated principal component scores. Upper left panel: scatterplot of the 1st versus 2nd PC scores for the column-specific term. Other panels: distribution of the first few PC scores for column-specific term, row-specific term and multiplicative term versus stimulation setting covariate.

Chapter 4

Multilevel matrix-variate analysis

Abstract

The number of studies where the primary measurement is a matrix is exploding. In response to this, we propose a statistical framework for modeling populations of repeated matrix-variate measurements. We use a linear mixed effect model to account for the multilevel design, while the 2D structure is handled via normal matrix-variate distribution. To achieve dimension reduction, we estimate and decompose the row- and column-specific covariance operators. The computational feasibility and performance of the approach is shown in extensive simulation studies. The method is motivated by and applied to a study that remotedly monitored physical activity of individuals diagnosed with congestive heart failure (CHF) over a 3- to 10-month period. Two primary goals of the study were: 1) to quantify and model the long-term patterns of physical activity in individuals with CHF; and 2) evaluate the possibility of predicting adverse health effects via continuous activity monitoring.

***keywords*: matrix-variate distribution, principal component analysis, mixed effect model**

4.1 Introduction

Modern studies often generate data in the format of matrices. For example, in studies measuring physical activity with fitness trackers and accelerometers studies where wearable devices are used to objectively record individual daily activity profiles, the data are often represented in the form of a matrix. For example, in the study by the Center for Advanced Cardiac Care at Columbia activity data is measured as activity counts in every minute for many months. To better conceptualize the data we denote by $Y_{ij}(d,t)$ the activity count for subject i at minute t of the day d within week j. Thus, for every subject/week pair the data are matrix variate of dimension 1440 (number of minutes in a day) by 7 (number of days in a week). To build up intuition, Figure 4.1 displays the log activity count profiles for two subjects (labeled Subjects 1 and 2) over 10 consecutive weeks.

Weeks are separated by a distinctive horizontal white stripe. The x-axis in both panels corresponds to minute of the day starting with 0 (midnight), while the y-axis corresponds to day from the beginning of monitoring. Darker red color corresponds to more intense activity, while light red and white correspond to low or no activity, respectively. Figure 1 provides striking visual information about the between- and within-subject variability. Indeed, Subject 1 exhibits higher levels of overall activity than Subject 2 (the left panel contains darker shades of red than the right panel). Moreover, the circadian patterns of activity of Subject 2 indicate a clear transition from night- to day-activity. In contrast, Subject 1 exhibits much larger day-to-day and within-day variability and has many nights when

the transition between sleep and wake periods is difficult to point out. These differences between the two subjects indicate the need for careful modeling of between- and within-subject variability. This multi-level study design together with the high-dimensionality of

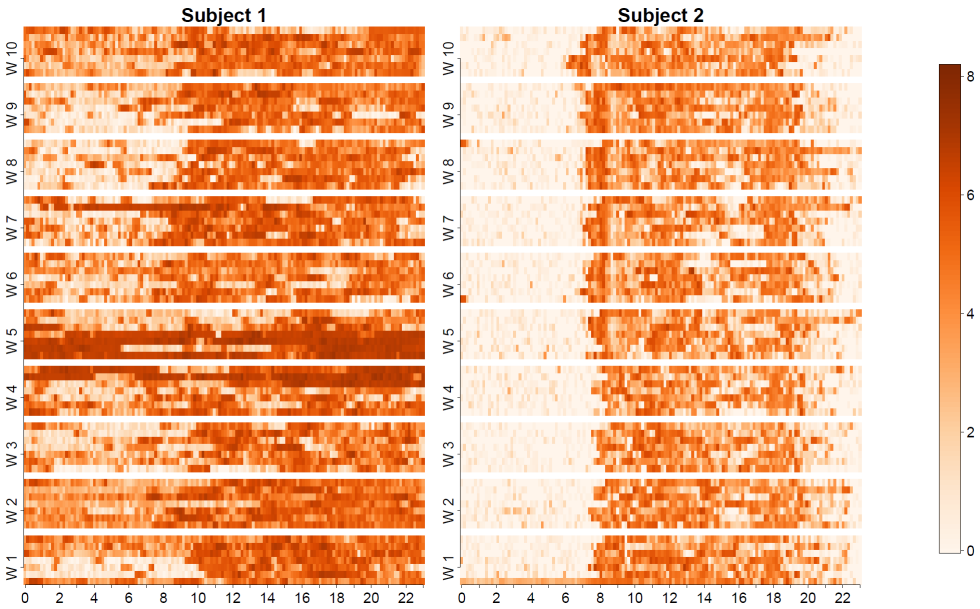

Figure 4.1: Log activity count plots for subject 1 and subject 2 across 10 weeks.

the data pose new analytic challenges. The primary goal of this paper is to address these challenges by proposing a multi-level principal component analysis for matrix-type data.

Principal component analysis (PCA) is a classic statistical method. It is often used for pattern estimation and dimensionality reduction. In recent years, there has been increasing interest in extending the simplicity and power of PCA to account for more complex data. To address the issue of multi-level study designs, Di et al. (2009); Greven et al. (2010); Di et al. (2014a); Zipunnikov et al. (2011); Shou et al. (2014) imposed a functional linear mixed effect model for the hierarchical study design. PCA was applied based on the co-

variance operators of the latent processes. Methods were extended to moderate- and high-dimensional matrix data using a rank reserving argument inspired by the singular value decomposition (SVD). To address the issue of matrix data structures, Caffo et al. (2010) and Crainiceanu et al. (2011) proposed the population value decomposition (PVD). PVD uses a two-stage singular value decomposition (SVD) to extract population-level principal components along each dimension.

Two-way matrix modelling is a highly active research area. Recently, Allen et al. (2014) introduced a two-way SVD for matrices to account for possibly correlated residuals. The correlation structure of the residuals was accounted for by using a separability assumption on the residuals. Allen (2013b); Huang et al. (2008); Witten et al. (2009); Lee et al. (2010); Tian et al. (2013); Zhang et al. (2013) looked at the regularization for two-way structured data, allowing sparsity or smoothness in row and column spaces. Along the line of regularization, Allen (2013a) extended methods to multi-way structures within the functional data framework using tensor decomposition. Spencer et al. (2001); Dien et al. (2003) proposed a two-stage PCA for matrix data, which performs a spatial PCA and a temporal PCA sequentially. Ye (2005) proposed a generalized low rank approximations of matrix (GLRAM) algorithm, which has lower computation time than SVD. Hung et al. (2012); Zhang and Zhou (2005) implemented a two-directional two-dimensional PCA under the separable covariance operator assumption. Zhou and Li (2014); Zhou et al. (2013) worked on supervised analysis for the matrix-variate data, proposing a framework for scalar-on-matrix regression. Hoff (2014) used separability assumptions on both the regression pa-

rameters and the covariance matrix to analyze relational data.

What we proposed here is related to these approaches. More specifically, we combine ideas of using explicit mixed effects modeling, assuming a separable covariance structure, and using PCA on the latent processes for implementing PCA on matrix data. The separability assumption reduces the dimensionality of the covariance operator from $O(D^2 \times T^2)$, where D and T is the number of rows and columns, respectively to $O(D + T)$. In our application, this is crucial as $D = 7$ and $T = 1440$. Thus, we propose a feasible method of moments algorithm for estimating and diagonalizing covariance operators. The explicit mixed effects model provides a direct way for modeling the multi-level design. The idea is different from PVD-type methods. Indeed, PVD provides a direct way for modeling the low-dimension score matrix shared by all levels of the data. The covariance decomposition proposed here accounts directly for the multi-level data structure. The follow-up PCA on each covariance component results in different PC subspace for different levels in the data.

The rest of the paper is organized as follows. In Section 4.2, we introduce our PCA model based on linear mixed effects model and separability assumption. We also propose the estimation algorithm in this section and extend our method to the data with white-noise. Section 4.3 provides a simulation study while Section 4.4 provides extensive results for the analysis on the activity study. We conclude the paper with a discussion of potential future research in Section 4.5. All technical proofs are delegated to Appendix.

4.2 Model and Estimation

Let $\mathbf{Y}_{ij}$ denote a matrix-variate observation for subject i at visit j of the dimension $D \times T$. For the accelerometry study, $\mathbf{Y}_{ij}$ is a 7×144-dimensional matrix that each row of $\mathbf{Y}_{ij}$ contains activity data for each day of the week.

We first introduce matrix-variate normal distribution. Matrix-variate random variable $\boldsymbol{Z}$ follows a matrix-variate normal distribution $\text{MN}_{D,T}(\boldsymbol{M}, \mathbf{C}, \mathbf{R})$, if $\text{vec}(\boldsymbol{Z})$ $\sim N_{DT}(\text{vec}(\boldsymbol{M}), \mathbf{R} \otimes \mathbf{C})$ (Dawid, 1981), where column- and row-specific covariance matrices $\mathbf{C}$ and $\mathbf{R}$ are such that $\mathbf{C} = E[(\boldsymbol{Z} - \boldsymbol{M})(\boldsymbol{Z} - \boldsymbol{M})^T]/\text{tr}(\mathbf{R})$ and $\mathbf{R} = E[(\boldsymbol{Z} - \boldsymbol{M})^T(\boldsymbol{Z} - \boldsymbol{M})]/\text{tr}(\mathbf{C})$, where $\text{tr}(\mathbf{C})$ and $\text{tr}(\mathbf{R})$ are the traces of $\mathbf{C}$ and $\mathbf{R}$. The probability density function for the random matrix $\boldsymbol{Z}$ has the form:

$$p(\boldsymbol{Z}|\boldsymbol{M}, \mathbf{C}, \mathbf{R}) = \frac{\exp\left(-\frac{1}{2}\text{tr}\left[\mathbf{R}^{-1}(\boldsymbol{Z} - \boldsymbol{M})^T\mathbf{C}^{-1}(\boldsymbol{Z} - \boldsymbol{M})\right]\right)}{(2\pi)^{DT/2}\|\mathbf{R}\|^{D/2}\|\mathbf{C}\|^{T/2}}, \tag{4.1}$$

To account for the nested design, we propose the following matrix-variate mixed effect model

$$\begin{cases} \mathbf{Y}_{ij} = \boldsymbol{M} + \mathbf{X}_i + \mathbf{W}_{ij}, i = 1, \ldots, I, j = 1, \ldots, J_i \\ \mathbf{X}_i \sim \text{MN}_{D,T}(\mathbf{0}, \mathbf{C}_X, \mathbf{R}_X) \\ \mathbf{W}_{ij} \sim \text{MN}_{D,T}(\mathbf{0}, \mathbf{C}_W, \mathbf{R}_W) \end{cases} \tag{4.2}$$

where $\boldsymbol{M}$ is the population mean and $\mathbf{X}_i$, $\mathbf{W}_{ij}$ are the subject-specific and subject-visit-specific deviations.

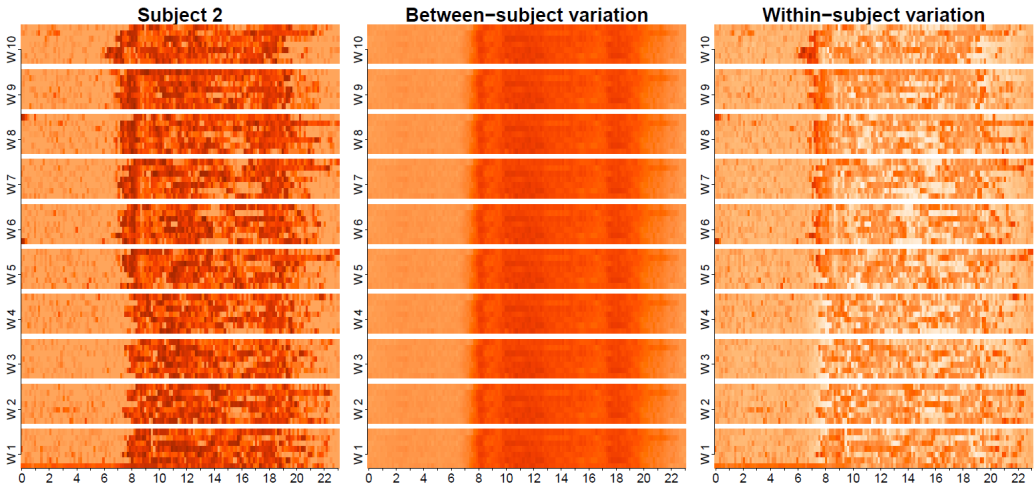

Figure 4.2: The figure illustrates the decomposition for a randomly chosen participant based on Equation (4.2). The left panel shows the deviation from the population mean $\mathbf{Y}_{ij} - \boldsymbol{M}$, the middle panel shows the estimated subject-specific deviation, $\mathbf{X}_i$, from the mean and the right panel shows the visit-specific deviation estimator, $\mathbf{W}_{ij}$, from the subject-specific mean.

To provide intuition behind our model, we refer to Figure 4.2 that gives a visual representation corresponding to (4.2) for a randomly chosen participant. The left panel depicts the residuals $Y_{ij} - \widehat{M}$, where $\widehat{M} = \frac{1}{J}\sum_{i,j} Y_{ij}$. The middle panel provides the estimator of $\widehat{X}_i = \frac{1}{J_i}\sum_i \{Y_{ij} - \widehat{M}\}$. Because $\widehat{X}_i$ is the same for every week, the middle panel contains the same subject-specific weekly profile. The right panel displays the following estimates of the visit specific deviations $\hat{W}_{ij} = Y_{ij} - \hat{M} - \hat{X}_i$. Thus, the population deviation, subject-specific deviations and visit-specific deviations are plotted in the left, middle and right panels respectively. Our primary goal is to use PCA to parsimoniously decompose the observed variability.

There are three steps in the proposed estimation algorithm: 1) obtain the within- and between-subject covariance matrices using fast method of moments approaches; 2) esti-

mate principal components by decomposing the within- and between-subject covariance matrices; 3) use a projection method to estimate the principal scores.

4.2.1 Covariance decomposition

In this section, we want to estimate $\mathbf{C}_W$, $\mathbf{R}_W$, $\mathbf{C}_X$ and $\mathbf{R}_X$ from $\boldsymbol{Y}_{ij}$'s. Shou et al. (2014) proposed a method-of-moments estimator for multi-level vector-type data. We adapt Shou et al. (2014) approach and use the separability property, which will allow us to conduct separate computations for each dimension. This will substantially reduce computational complexity.

More specifically, it is easy to show that

$$E\left[(\mathbf{Y}_{ij}-\mathbf{Y}_{kl})(\mathbf{Y}_{ij}-\mathbf{Y}_{kl})'\right]=\begin{cases}2\mathbf{C}_W\text{tr}(R_W), & i=k, j\neq l;\\ 2(\mathbf{C}_X\text{tr}(R_X)+\mathbf{C}_W\text{tr}(R_W)), & i\neq k.\end{cases} \tag{4.3}$$

and

$$E\left[(\mathbf{Y}_{ij}-\mathbf{Y}_{kl})'(\mathbf{Y}_{ij}-\mathbf{Y}_{kl})\right]=\begin{cases}2\mathbf{R}_W\text{tr}(C_W), & i=k, j\neq l;\\ 2(\mathbf{R}_X\text{tr}(C_X)+\mathbf{R}_W\text{tr}(C_W)), & i\neq k.\end{cases} \tag{4.4}$$

Let us denote $\boldsymbol{H}_W^C = 2\mathbf{C}_W\text{tr}(R_W)$, $\boldsymbol{H}_{XW}^C = 2(\mathbf{C}_X\text{tr}(R_X)+\mathbf{C}_W\text{tr}(R_W))$ and $\boldsymbol{H}_W^R = 2\mathbf{R}_W\text{tr}(C_W)$, $\boldsymbol{H}_{XW}^R = 2(\mathbf{R}_X\text{tr}(C_X)+\mathbf{R}_W\text{tr}(C_W))$. Let n_i be number of visits for subject i, $n=\sum_{i=1} n_i$ be the total number of weeks and $k=\sum_{i=1}^I n_i^2$ be the sum of squared

numbers of weeks. From Equation (4.3) and Equation (4.4), using a similar argument to the one in Shou et al. (2014), we propose the following method-of-moments estimators for $\boldsymbol{H}_W$ and $\boldsymbol{H}_{XW}$

$$\begin{cases} \widehat{\boldsymbol{H}}_W^C &= \frac{2}{k-n}\sum_{i=1}^{I}\sum_{j\neq l}(\mathbf{Y}_{ij}-\mathbf{Y}_{kl})(\mathbf{Y}_{ij}-\mathbf{Y}_{kl})' = \tilde{\boldsymbol{Y}}^C\mathbf{G}_W^C\tilde{\boldsymbol{Y}}^{C'} \\ \widehat{\boldsymbol{H}}_{XW}^C &= \frac{2}{n^2-k}\sum_{i\neq k}\sum_{j,l}(\mathbf{Y}_{ij}-\mathbf{Y}_{kl})(\mathbf{Y}_{ij}-\mathbf{Y}_{kl})' = \tilde{\boldsymbol{Y}}^C\mathbf{G}_{XW}^C\tilde{\boldsymbol{Y}}^{C'} \end{cases} \tag{4.5}$$

and

$$\begin{cases} \widehat{\boldsymbol{H}}_W^R &= \frac{2}{k-n}\sum_{i=1}^{I}\sum_{j\neq l}(\mathbf{Y}_{ij}-\mathbf{Y}_{kl})'(\mathbf{Y}_{ij}-\mathbf{Y}_{kl}) = \tilde{\boldsymbol{Y}}^R\mathbf{G}_W^R\tilde{\boldsymbol{Y}}^{R'} \\ \widehat{\boldsymbol{H}}_{XW}^R &= \frac{2}{n^2-k}\sum_{i\neq k}\sum_{j,l}(\mathbf{Y}_{ij}-\mathbf{Y}_{kl})'(\mathbf{Y}_{ij}-\mathbf{Y}_{kl}) = \tilde{\boldsymbol{Y}}^R\mathbf{G}_{XW}^R\tilde{\boldsymbol{Y}}^{R'} \end{cases} \tag{4.6}$$

where $\tilde{\boldsymbol{Y}}^C$ is a $D \times Tn$-dimensional matrix definded as [$\boldsymbol{Y}_{11}$,..., $\boldsymbol{Y}_{1n_1}$, $\boldsymbol{Y}_{21}$, ..., $\boldsymbol{Y}_{2n_2}$, ..., $\boldsymbol{Y}_{In_I}$]; $\tilde{\boldsymbol{Y}}^R$ is a $T \times Dn$-dimensional matrix defined as [$\boldsymbol{Y}_{11}'$, $\boldsymbol{Y}_{12}'$,..., $\boldsymbol{Y}_{1n_1}'$, $\boldsymbol{Y}_{21}'$, $\boldsymbol{Y}_{22}'$, ..., $\boldsymbol{Y}_{2n_2}'$, ..., $\boldsymbol{Y}_{In_I}'$]. Matrix $\mathbf{G}_W^C = \frac{2}{k-n}(\boldsymbol{D} - \boldsymbol{E}^T\boldsymbol{E})$ while matrix $\mathbf{G}_{XW}^C = \frac{2}{n^2-k}(n\boldsymbol{I}_n \otimes \boldsymbol{I}_D - (\mathbf{1}_n \otimes \boldsymbol{I}_D)(\mathbf{1}_n \otimes \boldsymbol{I}_D)' - \boldsymbol{D} + \boldsymbol{E}^T\boldsymbol{E})$ where $\boldsymbol{D} = \text{diag}\{\boldsymbol{N}_1 \otimes \boldsymbol{I}_D, \boldsymbol{N}_2 \otimes \boldsymbol{I}_D, ..., \boldsymbol{N}_I \otimes \boldsymbol{I}_D\}$ with $\boldsymbol{N}_i = n_i\boldsymbol{I}_{n_i}$; $\boldsymbol{E} = \text{diag}\{(\mathbf{1}_{n_1} \otimes \boldsymbol{I}_D)^T, (\mathbf{1}_{n_2} \otimes \boldsymbol{I}_D)^T, ..., (\mathbf{1}_{n_I} \otimes \boldsymbol{I}_D)^T\}$. Matrices $\mathbf{G}_W^R$ and $\mathbf{G}_{XW}^R$ have similar structures as matrices $\mathbf{G}_W^C$ and $\mathbf{G}_{XW}^C$.

4.2.2 Extension to more complex study designs

In addition to the two-level design we focus on, our approach elegantly extends to accommodate common nested and crossed designs. Koch (1967) provided a comprehensive

list of models for scalar data that emerge from various experimental designs. Shou et al. (2014) extended these designs to multivariate and functional data to capture a wide variety of correlation structures. Below, we describe how these designs can be fit for matrix-variate data through decomposition of the design specific column- and row-covariance matrices.

We consider a general m-way crossed model that can be expressed via a linear combination of latent matrix-variate variables as $\mathbf{Y}_{i_1 i_2 \cdots i_m} = \mathbf{M} + \mathbf{X}_{\mathcal{T}_1} + \mathbf{X}_{\mathcal{T}_2} + \cdots + \mathbf{X}_{\mathcal{T}_d}$. The d sub-index sets $\{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_d\}$ define the model design. For example, our two-way model from (4.2) corresponds to $d = 2$ with $\{\mathcal{T}_1, \mathcal{T}_2\} = \{i, ij\}$. Within this modeling framework latent terms $\mathbf{X}_{\mathcal{T}_i}$ are assumed to be zero-mean mutually independent matrix-variate normal random matrices. Consequently, the total column and row variability of matrix-variate observations can be decomposed into the term specific variabilities. Similar to (4.5) and (4.6), Shou et al. (2014) demonstrated that the covariance matrices of the latent processes $\mathbf{X}_{\mathcal{T}_i}$ can always be estimated via the "sandwich" estimator $\mathbf{C}_{X_{\mathcal{T}_i}} = \tilde{\mathbf{Y}}^C \mathbf{G}^C_{X_{\mathcal{T}_i}} \tilde{\mathbf{Y}}^{C\prime}$ and $\mathbf{R}_{X_{\mathcal{T}_i}} = \tilde{\mathbf{Y}}^R \mathbf{G}^R_{X_{\mathcal{T}_i}} \tilde{\mathbf{Y}}^{R\prime}$, where $\mathbf{G}_{X_{\mathcal{T}_i}}$'s are design specific matrices constructed according to the methods described in Section 4.2.1 and Shou et al. (2014).

4.2.3 Principal component estimation

For the estimation of principal components, we conduct matrix spectral decompositions on the estimated covariance matrix estimators for $\mathbf{X}_i$ and $\mathbf{W}_{ij}$.

For $\mathbf{X}_i$ we want to decompose $\text{Cov}[\text{vec}(X_i)]$ as $\Phi_X D_X \Phi_X^T$, where Φ_X is the orthog-

onal matrix and D_X is a diagonal matrix. Each column of Φ_X, $\phi_{\mathbf{k}}^{\mathbf{X}}$, is the estimated kth eigenfunction(principal component) evaluated at one combination of row- and columum-spaces. Similar to Section 4.2.1, directly diagonalizing the covariance matrix of vectorized $\boldsymbol{X}_i$ is computational intensive. Instead, using the separable covariance assumption, we can reduce the computational burden by conducting spectral decompositions on $\mathbf{C}_X$ and $\mathbf{R}_X$ separately. More specifically, we can first decompose $\mathbf{C}_X = \Phi_{C_X} D_{C_X} \Phi_{C_X}^T$ and $\mathbf{R}_X = \Phi_{R_X} D_{R_X} \Phi_{R_X}^T$ and make the following observation

$$\text{Cov}[\text{vec}(X_i)] = \mathbf{R} \otimes \mathbf{C} = [\Phi_{R_X} \otimes \Phi_{C_X}][D_{R_X} \otimes D_{C_X}][\Phi_{R_X} \otimes \Phi_{C_X}]^T.$$

Therefore, $\Phi_{R_X} \otimes \Phi_{C_X}$ is equivalent of Φ_X up to the permutation of columns, which provides the estimates of between-subject principal components. A similar implementation can be performed on the within-subject PCs for $\boldsymbol{W}_{ij}$.

4.2.4 Principal score estimation

Once the principal components are estimated, we express the observed data $\mathbf{Y}_{ij}$ as

$$\mathbf{Y}_{ij} = \boldsymbol{M} + \Phi_{C_X} \boldsymbol{\Gamma}_{\boldsymbol{i}}^{\boldsymbol{X}} \Phi_{R_X}^T + \Phi_{C_W} \boldsymbol{\Gamma}_{\boldsymbol{ij}}^{\boldsymbol{W}} \Phi_{R_W}^T, \tag{4.7}$$

where $\boldsymbol{\Gamma}_{\boldsymbol{i}}^{\boldsymbol{X}} \sim \text{MN}_{D,T}(0, D_{C_X}, D_{R_X})$ and $\boldsymbol{\Gamma}_{\boldsymbol{ij}}^{\boldsymbol{W}} \sim \text{MN}_{D,T}(0, D_{C_W}, D_{R_W})$, where D_{C_X}, D_{R_X}, D_{C_W}, D_{R_W} are the diagonal matrices of eigenvalues estimated in Section 4.2.3.

For inference and estimation purpose, we further assume that $\boldsymbol{\Gamma_i^X}$ and $\boldsymbol{\Gamma_{ij}^W}$ have a matrix-variate normal distribution (MN). Our goal is to estimate the scores for the first $K_1 \times K_2$ PCs of $\boldsymbol{X}_i$ and the first $L_1 \times L_2$ PCs of $\boldsymbol{W}_{ij}$.

Given $\mathbf{Y}_{ij}$, the scores can be estimated directly based on Equation (4.7). By vectorizing Equation (4.7), we have

$$\begin{cases} \text{vec}(\mathbf{Y}_{ij}) = \text{vec}(\mathbf{M}) + (\boldsymbol{\Phi}_{R_X} \otimes \boldsymbol{\Phi}_{C_X})\text{vec}(\boldsymbol{\Gamma}_i^X) + (\boldsymbol{\Phi}_{R_W} \otimes \boldsymbol{\Phi}_{C_W})\text{vec}(\boldsymbol{\Gamma}_{ij}^W) \\ \text{vec}(\boldsymbol{\Gamma}_i^X) \sim N(0, \mathbf{D}_{R_X} \otimes \mathbf{D}_{C_X}) \\ \text{vec}(\boldsymbol{\Gamma}_{ij}^W) \sim N(0, \mathbf{D}_{R_W} \otimes \mathbf{D}_{C_W}) \end{cases} \tag{4.8}$$

Equation (4.8) is a linear mixed model with the random effects $\text{vec}(\boldsymbol{\Gamma}_i^X)$ and $\text{vec}(\boldsymbol{\Gamma}_{ij}^W)$ being the estimands of interest. Thus, the mixed model inferential machinery can be used to estimate the scores using the best linear unbiased prediction (BLUP). The detailed BLUP derivation is presented in Appendix A1.

When the dimensionality $D \times T$ grows large, the implementation of above method may not be possible due to computational memory constrain. Instead we can use an approximation method which is presented in Appendix A1.

4.2.5 Data with white noise

Our model assumes that matrix-variate observations do not containt measurement error or that noise is negligible and can be incorporated into a smooth covariance structures.

However, our methods can easily be extended to data contaminated by white noise $\boldsymbol{E_{ij}} \sim MN(0, \boldsymbol{I_D}, \boldsymbol{I_T})$, i.e.,

$$\mathbf{Y}_{ij} = \boldsymbol{M} + \mathbf{X}_i + \mathbf{W}_{ij} + \boldsymbol{E_{ij}}, \tag{4.9}$$

To ensure identifiability, we need to assume that at least one dimension of $\mathbf{Y}_{ij}$ is smooth. Suppose, for example, that the process is a smooth function in the row space, then $\mathbf{W}_{ij}$ and $\boldsymbol{E_{ij}}$ can be separated by smoothing the off-diagonal surface of $\widehat{\mathrm{C}}_W$ (Staniswalis and Lee, 1998).

However, sometimes this smoothing techniques will become computational infeasible if all dimensions are very large. In this case, the methods in Shabalin and Nobel (2013) could be considered as a powerful and feasible alternative.

4.3 Simulation

To better understand the multi-level separable model and the performance of our proposed algorithm, we conduct following simulation studies.

We simulated the data based on the following model,

$$\begin{aligned} y_{ij}(s,t) = &\sum_{k_{C_X}=1}^{4} \sum_{k_{R_X}=1}^{3} \xi^X_{ik_{C_X}k_{R_X}} \phi^{C_X}_{k_{C_X}}(s) \psi^{R_X}_{k_{R_X}}(t) \\ &+ \sum_{k_{C_W}=1}^{2} \sum_{k_{R_W}=1}^{2} \xi^W_{ijk_{C_W}k_{R_W}} \phi^{C_W}_{k_{C_W}}(s) \psi^{R_W}_{k_{R_W}}(t) + \varepsilon_{ij}(s,t) \end{aligned}$$

where $\xi^X_{ik_{C_X}k_{R_X}}$ and $\xi^W_{ijk_{C_W}k_{R_W}}$ are the subject-specific and visit-specific scores while $\phi^{C_X}_{k_{C_X}}(s)$, $\phi^{C_W}_{k_{C_W}}(s)$ and $\psi^{R_X}_{k_{R_X}}(t)$, $\psi^{R_W}_{k_{R_W}}(t)$ are the corresponding eigenfunctions along the row and column domains. The simulated $\mathbf{Y_{ij}}$ is given by $y_{ij}(s,t)$ evaluated on an equally-spaced 30 by 20 grid at $[0,1] \times [0,1]$.

True eigenfunctions are

$$\phi^{C_X}_1(s) = \sqrt{2}\sin(2\pi s) \quad \phi^{C_X}_2(s) = \sqrt{2}\cos(2\pi s) \quad \phi^{C_X}_3(s) = \sqrt{2}\sin(4\pi s)$$
$$\phi^{C_X}_4(s) = \sqrt{2}\cos(4\pi s)$$
$$\psi^{R_X}_1(t) = \sqrt{2}\sin(6\pi t) \quad \psi^{R_X}_2(t) = \sqrt{2}\cos(6\pi t) \quad \psi^{R_X}_3(t) = \sqrt{2}\sin(8\pi t)$$
$$\phi^{C_W}_1(s) = \sqrt{2}\sin(2\pi l) \quad \phi^{C_W}_2(s) = \sqrt{2}\cos(2\pi s) \quad \psi^{R_W}_1(t) = \sqrt{2}\sin(6\pi t)$$
$$\psi^{R_W}_2(t) = \sqrt{2}\cos(6\pi t)$$

and the subject-specific scores are generated from the matrix normal distribution $MN(0,$ diag$\{1, 0.5, 0.25\}$, diag$\{1, 0.5, 0.25, 0.125\})$ while the visit-specific scores are generated from $MN(0, \text{diag}\{1, 0.5\}, \text{diag}\{1, 0.5\})$. The white noise is generated from a normal distribution $N(0, \sigma^2_\varepsilon)$ where σ^2_ε is determined by the level of signal-to-noise ratio. In this simulation study, we vary the signal-to-noise ratio into four levels: $+\infty, 10, 1, 0.1$, where the $+\infty$ corresponds to the case without white noise. The number of subjects $I = 300$ and number of visits $J = 2$.

For each simulation settings, we generated 100 datasets. Figures A1 through A4 in the appendix display fitted principal components (grey lines) versus true principal components (black lines) for the four different signal-to-noise ratios. In general, there is very

good visual agreement between the true and estimated curves and the quality of estimation increases with the signa-to-noise ratio. To further quantify the performance, Table 4.1 displays average MSE between the estimated and the true eigenvectors using 100 simulation data. As observed in the Figures A1-A4 in the appendix, we see that the MSE increases as the signal to noise decreases. This is expected and reassuring, which supports the idea that the proposed approach is numerically stable and produces reproducible results even in very low signal-to-noise scenarios. Table 4.2 displays the average MSE between estimated and true scores.

Signal-to-noise	$+\infty$	10	1	0.1
$\phi_1^{C_X}(s)$	0.0246	0.0228	0.0232	0.0312
$\phi_2^{C_X}(s)$	0.0384	0.0386	0.0399	0.0652
$\phi_3^{C_X}(s)$	0.0336	0.0324	0.0349	0.1022
$\phi_4^{C_X}(s)$	0.0178	0.0176	0.0217	0.2300
$\psi_1^{R_X}(t)$	0.0177	0.0162	0.0165	0.0224
$\psi_2^{R_X}(t)$	0.0283	0.0270	0.0279	0.0459
$\psi_3^{R_X}(t)$	0.0174	0.0186	0.0196	0.0641
$\phi_1^{C_W}(s)$	0.0080	0.0131	0.0137	0.0286
$\phi_2^{C_W}(s)$	0.0080	0.0131	0.0145	0.0573
$\psi_1^{R_W}(t)$	0.0107	0.0102	0.0106	0.0229
$\psi_2^{R_W}(t)$	0.0107	0.0102	0.0113	0.0519

Table 4.1: Average MSE between estimated and true eigenvectors using 100 simulation data.

Signal-to-noise	$+\infty$	10	1	0.1
Between-subject score	0.0835	0.0836	0.0868	0.1192
Within-subject score	0.2219	0.2248	0.2272	0.2514

Table 4.2: Average MSE between estimated and true scores using 100 simulation data.

4.4 Data application

Fifty nine patients of the Advanced Cardiac Care Center of Columbia University Medical Center diagnosed with congestive heart failure (CHF) wore Actical, an accelerometer device that continuously recorded physical activity over a three to nine months period. Over the course of the study, twenty four subjects were either hospitalized or had an emergency room visit. The aim of the study is to explore physical activity (PA) in this clinical population in real-life context, understand the main patterns of day-to-day and week-to-week variability, and identify possible associations between patterns of PA and adverse clinical events. Subjects with less than 10 weeks data were excluded in this data analysis resulting in a sample of 51 subjects.

Minute level activity counts were log- transformed, $y = \log(x+1)$, to reduce the strong skewness characteristic typically observed in PA data. Last, we averaged the transformed log counts into within 10-minute non-overlapping time windows. Because we expect major within-week variability associated especially with week-ends, we treat the measurements of one week as one observation unit. Therefore, each observation unit will be stored in a 7×144 matrix. Every subject has between 13 and 47 observed weeks. The left panel in Figure 4.3 displays the average week acitivity across all subject and all visits. The middle panel display the population mean as a function of time for every day of the week (Monday through Sunday). The right panel in Figure 4.3 displays the average activity as a function of day of the week at 6 time intervals during the day. Figure 4.3 indicates that activity

is reduced during the weekends with a more pronounced reduction on Sunday than on Saturday. Also, the reduction in activity is more pronounced at 8am over the week-end, while activity around 12am increases slightly during the week-end.

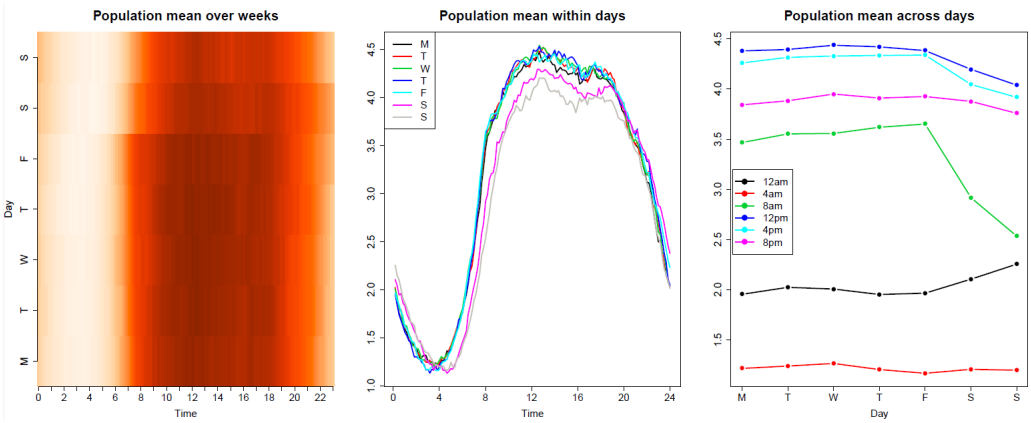

Figure 4.3: Population mean of log activity counts within days and across days.

4.4.1 Eigenvalues and eigenfunctions

In this section, we apply our proposed multilevel matrix PCA model to the data. Table 4.3 displays the estimated eigenvalues and related information for the first six components for $\mathbf{C}_X$, $\mathbf{R}_X$, $\mathbf{C}_W$ and $\mathbf{R}_W$.

One question that may be asked is how much variability each level (within- and between-subject levels) contributes. To answer this question, we introduce a parameter ρ which is defined as the proportion of variability explained by the between-subject level. Note that ρ generalizes the intra-class correlation (ICC) for scalars. Since $\boldsymbol{Y}_{ij}$ is a matrix, we use the matrix trace operator to generalize ICC in the matrix space. From Equation

Time between-subject components (R_X)						
Component	1	2	3	4	5	6
Eigen Val	12.82	7.89	3.47	1.33	1.04	0.78
Proportion Var	0.43	0.26	0.12	0.04	0.03	0.03
Cum proportion	0.43	0.69	0.81	0.85	0.89	0.91
Day between-subject components (C_X)						
Component	1	2	3	4	5	6
Eigen Val	26.92	1.60	0.41	0.32	0.18	0.15
Proportion Var	0.91	0.05	0.01	0.01	0.01	0.01
Cum proportion	0.91	0.96	0.98	0.99	0.99	1.00
Time within-subject components (R_W)						
Component	1	2	3	4	5	6
Eigen Val	10.07	3.49	3.14	2.14	1.85	1.55
Proportion Var	0.19	0.06	0.06	0.04	0.03	0.03
Cum proportion	0.19	0.25	0.31	0.35	0.38	0.41
Day within-subject components (C_W)						
Component	1	2	3	4	5	6
Eigen Val	14.81	7.82	6.80	6.50	6.23	5.85
Proportion Var	0.28	0.15	0.13	0.12	0.12	0.11
Cum proportion	0.28	0.42	0.55	0.67	0.78	0.89

Table 4.3: Eigenvalues for PCA for accelerometry study

(4.2), we define

$$\begin{aligned}\rho_X &= \frac{\text{var}\left\{\text{vec}(\boldsymbol{X}_i)\right\}}{\text{var}\left\{\text{vec}(\boldsymbol{X}_i)\right\} + \text{var}\left\{\text{vec}(\boldsymbol{W}_{ij})\right\}} \\ &= \frac{\text{tr}(\mathbf{C}_X)\text{tr}(\mathbf{R}_X)}{\text{tr}(\mathbf{C}_X)\text{tr}(\mathbf{R}_X) + \text{tr}(\mathbf{C}_W)\text{tr}(\mathbf{R}_W)} \\ &= \frac{\text{tr}(\boldsymbol{D}_{C_X})\text{tr}(\boldsymbol{D}_{R_X})}{\text{tr}(\boldsymbol{D}_{C_X})\text{tr}(\boldsymbol{D}_{R_X}) + \text{tr}(\boldsymbol{D}_{C_W})\text{tr}(\boldsymbol{D}_{R_W})}.\end{aligned}$$

In this study, we estimate $\hat{\rho} = 0.234$, that is, 23.4% of variability in weekly physical activity is attributable to the between-subject level variability. An alternative and more intuitive interpretation is to treat ρ_X as the correlation between two randomly chosen weeks from

the same subject.

Next we investigate the variability of $\mathbf{C}$ and $\mathbf{R}$ within each level. A characteristic of the variability for both within-subject components is not particularly well concentrated. Indeed, for $\mathbf{R}_W$, the first 6 components explain only 41% of variability. However, for the between-subject components, the variability is much more concentrated on the leading principal components. For example, the first four PCs of $\mathbf{R}_X$ explain over 85% of the variation.

Figure 4.4 displays the first four principal components adjusted by the corresponding eigenvalues for $\mathbf{C}_X$, $\mathbf{R}_X$, $\mathbf{C}_W$ and $\mathbf{R}_W$ (i.e. $\sqrt{\lambda}_k\phi_k$). For $\mathbf{C}_X$ (Day-Between), the first principal component (black) can be interpreted as the overall day deviation. i.e. some people are less active throughout the week. The second principal component (red) corresponds to the contrast between higher (lower) activity on weekdays versus lower(higher) activity on weekends. The third component (green) is similar to the second component but contains a contrast between Saturday and Sunday. The fourth component (blue) reflects a fluctuating activity pattern during the week. The principal components for $\mathbf{C}_W$ (Day-Within) are similar as $\mathbf{C}_X$ and have a similar interpretation.

For $\mathbf{R}_X$ (Time-Between), the first principal component (black) can be interpreted as the overall deviation of activity across times of the day. The second component (red) has a global maximum at 8am, which is the usual time of waking up with a marked decrease after 8am. The third component (green) corresponds to the contrast between activity during sleeping/night hours (11pm - 7am) and awake/day times (8am to 9pm). The fourth compo-

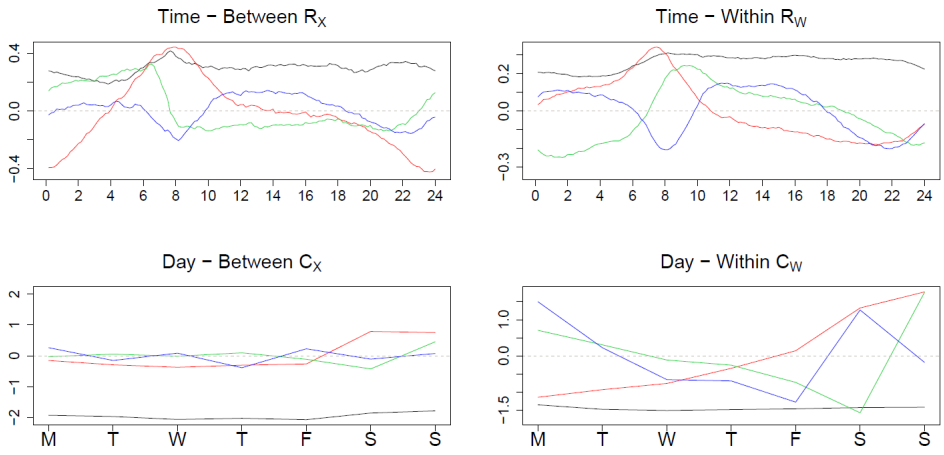

Figure 4.4: Principal components for $\mathbf{C}_X$, $\mathbf{R}_X$, $\mathbf{C}_W$ and $\mathbf{R}_W$, adjusted by eigenvalues. Black, red, green, blue lines are the 1st, 2nd, 3rd, and 4th principal components, respectively.

nent (blue) corresponds to the localized activity in the morning (8am) in contrast with the rest of the day. The fourth principal component is different from the second, because the second component has much larger fluctuations between maximum and minimum peaks of activity and much more abrupt transitions between peak, moderate, and low activity. The principal components for $\mathbf{R}_W$ (Time-Within) have a similar shape and interpretation but are much smoother because each subject has many weeks (repetitions).

4.4.2 Score matrices and covariates

During the study, twenty four participants encountered various events such as hospitalization (9), emergency room visit (8), intercurrent illness (4) and outpatient procedure (3).

Using the estimation method in Section 4.2.4, we derived the between- and within-

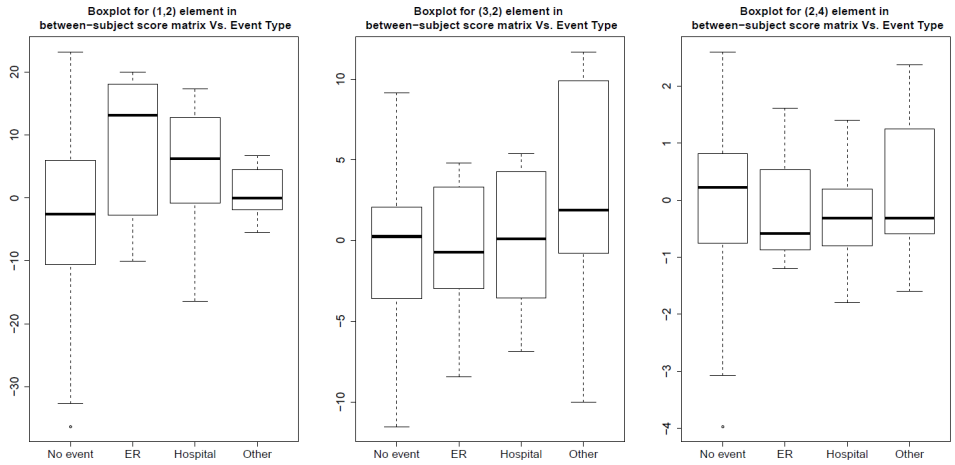

Figure 4.5: Boxplots of elements in between-subject score matrices grouped by event types.

subject score matrices and investigated the association between score matrices and the event-type covariate. We combine intercurrent illness and outpatient procedures into one event type and label it as "Other". Figure 4.5 displays the boxplots of three selected elements in between-subject score matrices grouped by event types. For example, in the upper panel of Figure 4.5, the (1,2) element stands for the score of the first principal component for the time dimension and the second principal component on the day dimension. The subjects with all four events tend to have higher scores than those without events. The interpretation is that for (3,2) elements and (2,4) elements, the between-subject scores are higher for subjects with intercurrent illness and outpatient procedures, respectively.

Figure 4.6 displays the principal component matrices corresponding to the scores in Figure 4.5. Principal component matrices are defined as the outer products between principal components for the time and day dimension. For example, the upper left panel of Figure 4.6 displays the principal matrix generated by the first principal component for the

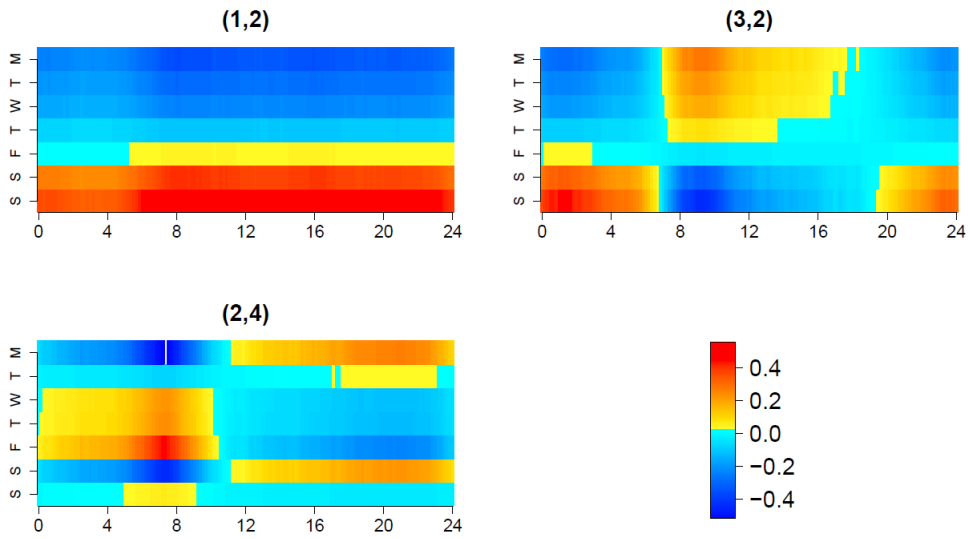

Figure 4.6: Principal matrices which are the outer products between different principal components for the time and day dimension.

time dimension and the second principal component in the day dimension, which can be interpreted as the contrast between weekdays and weekends. The principal matrix (3,2) characterizes the interaction between the weekday-weekend effects and sleep-wake effects.

Next, we continue to investigate the association between event type and between-subject scores while adjusting for other covariates. To simplify the problem, we classify event types "Emergency room", "Hospitalization", "Intercurrent illness" and "Outpatient procedures" into one class "With event". We consider the following logistic regression model

$$\text{logit}\{\Pr(Y_i = 1)\} = \beta_0 + \beta_1 \xi_i^{12} + \beta_2 \xi_i^{32} + \beta_3 \xi_i^{24} + R_i^T \gamma \tag{4.10}$$

where Y_i is the binary outcome for the presence of the event, ξ_i are (1,2), (3,2) and (2,4)

	Model 1	Model 2	Model 3	Model 4
Score (1,2)	0.012(0.005)*	0.011(0.005)*	0.010(0.005)*	0.011(0.035)*
Score (3,2)	0.016(0.012)	0.015(0.013)	0.015(0.013)	0.013(0.013)
Score (2,4)	0.024(0.054)	0.014(0.056)	0.026(0.057)	0.029(0.057)
Sex		0.113(0.140)	0.099(0.141)	0.056(0.147)
Age			0.004(0.004)	0.005(0.005)
BMI				0.012(0.011)

Table 4.4: Models for association between events and between-subject scores. For the variable sex, female is the reference group and an asterisk indicates significance at level 0.05

elements in between-subject score matrix for subject i and R_i is a vector of covariates containing subject age, sex, and BMI. The regression results are summarized in Table 4.4. Models included combinations of covariates including sex, age and body mass index (BMI). All four models indicated that the (1,2) element in the score matrix (the score for the first principal component in the time dimension and the second principal component in the day dimension) is strongly and positively associated with whether the subject will have an event. The magnitude of association varies slightly with the amount of covariates adjustment. For example, Model 3 estimates that a subject with one unit higher in the (1,2) score has $e^{0.010} = 1.01$ times the odds of the event occur, controlling for sex and age. Considering the scale of the score (the (1,2) PC scores have mean zero and standard deviation 4.53), standardized coefficients are easier to interpret. After standardizing, one standard deviation increase in the (1,2) PC score is associated with an odds ratio $e^{0.453} = 1.57$. Model 4, which adjusts for all the covariates, estimated an odds ratio of $e^{0.11} = 1.116$ per unit increase in the (1,2) PC score, or an odds ratio $e^{0.4983} = 1.64$ per one standard deviation increase in the (1,2) PC score. The (3,2) and (2,4) principal components were not found to

be associated with event occurrence after adjusting for the (1,2) component effect..

Last, we investigated the within-subject scores and their association with the events. Specifically, we are looking at subjects with "Hospitalization" and "Emergency Room". In Figure 4.7, we compare how scores are distributed in the first week, during the event and after the event. Similar to the interpretation for the between-subject score matrix, the (i, j) entry in the within-subject score matrix is the score for the ith principal component in the time dimension and jth principal component in day dimension. Each line denotes one trajectory from a single subject. Brown lines stands for Hospitalization (9 total events) while green lines stands for Emergency room visit (total of 8 events). Both scores reveal a similar pattern, which corresponds to the type of behavior expected for measurements that are sensitive to health events. Scores are lower in the first week and go up during the event week. One week after the event, the scores are still higher but get closer to the pre-event level. This inverse-U shape pattern corresponds to the expected recovery pattern of patients with hospitalization and outpatient procedures.

4.5 Discussion

This article introduces a multi-level matrix principal component analysis model, using separable covariance matrix assumptions and linear mixed effect modeling. We proposed method-of-moments estimators for principal components and developed two algorithms to estimate the principal scores. When we applied the method to data from the accelerometry

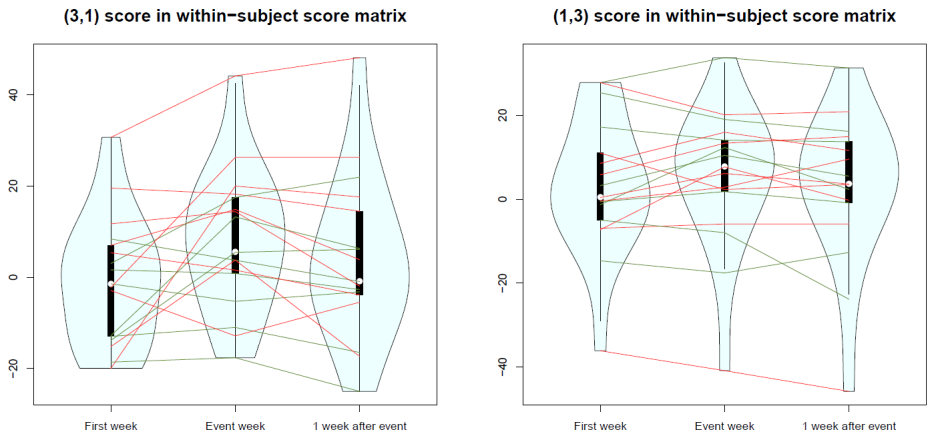

Figure 4.7: Violin plots for selected scores in within-subject score matrices in the first week, during event and after event.

study, we estimated several relevant patterns inherent in the data and quantified the amount of variability captured by the each level. We also studied the relationship between score matrices and health outcomes occurrence and timing.

Our approach also suggests several future directions of research. First, the estimated covariance matrix cannot be guaranteed to be positive definite when the number of subjects is less than the maximum of the dimensions of rows and columns. One potential solution is to work on some sub-model of the proposed separable models. For example, Fosdick and Hoff (2014) decomposed the covariance matrix as the sum of a reduced-rank matrix and a diagonal matrix and then approximated the reduced-rank matrix by a factor model using MLE. Second, a goodness-of-fit test should be developed to test whether the separable covariance structure assumption holds. Related work include Mitchell et al. (2006); Lu and Zimmerman (2005) where likelihood ratio tests were proposed. However, even if the separability assumption fails, the method proposed here is still useful for exploratory

purpose. Another alternative is to use population value decomposition (Crainiceanu et al., 2011) after multilevel variation decomposition step because our model is a special case of PVD. Last but not least, similar as Allen (2013a), we can extend our method to deal with multi-way structured data.

Chapter 5

Discussion and Future Work

While the scientific applications are quite different, the methods proposed here address three important challenges associated with modern data: high dimensionality, complexity and size. The key questions that we addressed are 1) does the structure influence the type of analysis? 2) if it does, then what are reasonable and feasible strategies to incorporate data structure into the analysis? and 3) can the specific data structures guide the dimensionality reduction procedures? To address these questions, we proposed and implemented three general methods motivated by specific scientific applications focused on structural connectivity (DTI), functional connectivity (fMRI), and activity monitoring (accelerometry studies).

The first method proposed a novel linear regression approach for analyzing the relationship between cognitive disability and white matter integrity as measured by fractional

anisotropy obtained from DTI-MRI studies. we proposed a Bayesian regression model with a latent binary indicator that controls whether or not there is an effect at each voxel. We use an Ising prior, which favors sparsity and spatial contiguity. Methods were applied to a study of association between anatomic connectivity (as measured by fractional anisotropy) and cognitive outcomes (as measured by the PASAT scores). We focused on subjects with Multiple Sclerosis. Results show that the most of the predictive regions are located at the corpus callosum, which agrees with previously published work (Barnea-Goraly et al., 2004; Keller et al., 2007). Our methodological approaches provide more in-depth analysis in a framework that accounts for the potential confounding effects of the other voxels in the brain.

The methodology has several limitations. First, if the Ising-prior hyper-parameters are estimated by cross-validation, the computation time is high; this can be partially alleviated by parallel computing. Alternatively, a pilot cross-validation study could be performed on a subregion of the brain and the estimated parameters can then be applied to the entire image. Second, our approach is a hybrid between a Bayesian and a frequentist approach, where the hyper-parameter are estimated via cross-validation. A fully Bayesian approach might provide a more integrated and philosophically satisfying alternative. Third, in some analyses, one may be interested in smoothing the non-zero regression parameters using a CAR prior (Goldsmith et al., 2013).

Motivated by studies that collect matrix-valued data, separable two-way matrix-variate models are proposed using explicit latent process modeling. Identifiability conditions are

introduced and method-of-moments estimators are provided for the covariance matrices of all latent processes. Principal component analysis is then used for dimensionality reduction at the level of individual spatial and temporal processes. Methods can be applied to data observed with or without white noise. Methods were applied to the data from the fMRI study, we identified important patterns inherent in the data and quantified the amount of variability captured by the various components.

Our work suggest multiple future directions of research. First, in our spatio-temporal decompositions, the estimated covariance matrix cannot be guaranteed to be positive definite when the number of subjects is smaller than the maximum of space and time dimensions. One idea could be to work on some sub-model of the proposed separable models. For example, Fosdick and Hoff (2014) decomposed the covariance matrix to the sum of a reduced-rank matrix and a diagonal matrix and then approximated the reduced-rank matrix by a factor model using MLE. Second, the noise-free version of our method of moments estimators are similar to the two-directional two-dimensional PCA (Zhang and Zhou, 2005), and may be affected by white noise. This potential problem could be addressed by using the off-diagonal smoothing technique proposed by Staniswalis and Lee (1998). Another possiblity is to consider the multilinear estimator described in Hung et al. (2012) and use iterative alternating least squares estimation. Third, we did not take the multi-level data structure into account. To solve this problem, we can implement decomposition ideas proposed by Shou et al. (2014) before separating the spatial and temporal variability spatio-temporal variations are separated. Fourth, methods could be adapted to sparse functional observa-

tions (Di et al., 2014b). Last, there is a need for a rigurous hypothesis testing framework for the various assumptions of separability. However, even if models do not hold, they can still be very useful for exploratory purposes.

The third method introduces a multi-level matrix principal component analysis model, using separable covariance matrix assumptions and linear mixed effect modeling. We proposed method-of-moments estimators for principal components and developed two algorithms to estimate the principal component scores. When we applied the method to data from the accelerometry study, we estimated several relevant patterns inherent in the data and quantified the amount of variability captured by each level. We also studied the relationship between score matrices and health outcomes occurrence and timing.

This approach also suggests several future directions of research. Similar to the previous method, we could try to identify estimation approaches that are guaranteed to provide positive definite covariance estimators. Also, testing for separability of the covariance structure would be necessary. Moreover, the approach could be extended to deal with multi-way structured data (Allen, 2013a).

A1 Appendix to Chapter 4

Finding principal scores using BLUPs under full model

Here, we provide details of calculating principal scores in model (4.8). We assume noise-free scenario under full model method. We follow Zipunnikov et al. (2011) and write model (4.8) as $\text{vec}(\mathbf{Y}_i) = \mathbf{A}z_i$, where $\text{vec}(\boldsymbol{Y_i}) = \big(\text{vec}(\boldsymbol{Y_{i1}})^T, \text{vec}(\boldsymbol{Y_{i2}})^T, ..., \text{vec}(\boldsymbol{Y_{iJ}})^T\big)^T$, $\boldsymbol{A} = (\mathbf{1_J} \otimes \boldsymbol{\Phi_{R_X}} \otimes \boldsymbol{\Phi_{C_X}}, \boldsymbol{I_J} \otimes \boldsymbol{\Phi_{R_W}} \otimes \boldsymbol{\Phi_{C_W}})$ and $z_i = \big(\text{vec}(\boldsymbol{\Gamma_i^X})^T, \text{vec}(\boldsymbol{\Gamma_{i1}^W})^T, \text{vec}(\boldsymbol{\Gamma_{i2}^W})^T, ..., \text{vec}(\boldsymbol{\Gamma_{iJ}^W})^T\big)^T$. So the BLUP of z_i would be

$$
\begin{aligned}
\hat{z}_i &= (\boldsymbol{A^T A})^{-1} \boldsymbol{A^T} \text{vec}(\boldsymbol{Y_i}) \\
&= \begin{pmatrix} {}^J\boldsymbol{I}_{\boldsymbol{N_{C_X} N_{R_X}}} & \mathbf{1}_{\boldsymbol{J}}^T \otimes \boldsymbol{\Phi}_{\boldsymbol{R_X}}^{\boldsymbol{T}} \boldsymbol{\Phi}_{\boldsymbol{R_W}} \otimes \boldsymbol{\Phi}_{\boldsymbol{C_X}}^{\boldsymbol{T}} \boldsymbol{\Phi}_{\boldsymbol{C_W}} \\ \mathbf{1_J} \otimes \boldsymbol{\Phi}_{\boldsymbol{R_W}}^{\boldsymbol{T}} \boldsymbol{\Phi}_{\boldsymbol{R_X}} \otimes \boldsymbol{\Phi}_{\boldsymbol{C_W}}^{\boldsymbol{T}} \boldsymbol{\Phi}_{\boldsymbol{C_X}} & \boldsymbol{I}_{\boldsymbol{JN_{C_W} N_{R_W}}} \end{pmatrix}^{-1} \\
&\times \begin{pmatrix} \text{vec}\left[(\boldsymbol{\Phi}_{\boldsymbol{R_X}}^T \otimes \boldsymbol{\Phi}_{\boldsymbol{C_X}}^T) \boldsymbol{Y}_i \mathbf{1}_J\right] \\ \text{vec}\left[(\boldsymbol{\Phi}_{\boldsymbol{R_W}}^T \otimes \boldsymbol{\Phi}_{\boldsymbol{C_W}}^T) \boldsymbol{Y}_i\right] \end{pmatrix}
\end{aligned}
$$

By block inversion, we can further simplify

$$
(\boldsymbol{A^T A})^{-1} = \begin{pmatrix} \boldsymbol{C_{11}} & \boldsymbol{C_{12}} \\ \boldsymbol{C_{21}} & \boldsymbol{C_{22}} \end{pmatrix}
$$

where

$$
\begin{aligned}
\boldsymbol{C_{11}} &= \left(J\boldsymbol{I}_{\boldsymbol{N_{C_X}N_{R_X}}} - J\boldsymbol{\Phi}_{\boldsymbol{R_X}}^{\boldsymbol{T}}\boldsymbol{\Phi}_{\boldsymbol{R_W}}\boldsymbol{\Phi}_{\boldsymbol{R_W}}^{\boldsymbol{T}}\boldsymbol{\Phi}_{\boldsymbol{R_X}} \otimes \boldsymbol{\Phi}_{\boldsymbol{C_X}}^{\boldsymbol{T}}\boldsymbol{\Phi}_{\boldsymbol{C_W}}\boldsymbol{\Phi}_{\boldsymbol{C_W}}^{\boldsymbol{T}}\boldsymbol{\Phi}_{\boldsymbol{C_X}} \right)^{-1} \\
&= \frac{1}{J}\left(\boldsymbol{\Phi}_{\boldsymbol{R_X}}^{\boldsymbol{T}} \otimes \boldsymbol{\Phi}_{\boldsymbol{C_X}}^{\boldsymbol{T}} \left\{ \boldsymbol{I} - (\Phi_{R_W}\Phi_{R_W}^T \otimes \Phi_{C_W}\Phi_{C_W}^T) \right\}^{-1} \boldsymbol{\Phi}_{\boldsymbol{R_X}} \otimes \boldsymbol{\Phi}_{\boldsymbol{C_X}} \right) \\
\boldsymbol{C_{12}} &= -\boldsymbol{C_{11}}\left(\mathbf{1}_{\boldsymbol{J}}^T \otimes \boldsymbol{\Phi}_{\boldsymbol{R_X}}^{\boldsymbol{T}}\boldsymbol{\Phi}_{\boldsymbol{R_W}} \otimes \boldsymbol{\Phi}_{\boldsymbol{C_X}}^{\boldsymbol{T}}\boldsymbol{\Phi}_{\boldsymbol{C_W}} \right) \\
\boldsymbol{C_{22}} &= \left(\boldsymbol{I}_{\boldsymbol{N_{C_W}N_{R_W}}} - (\mathbf{1}_{\boldsymbol{J}}\mathbf{1}_{\boldsymbol{J}}^{\boldsymbol{T}}) \otimes \boldsymbol{\Phi}_{\boldsymbol{R_W}}^{\boldsymbol{T}}\boldsymbol{\Phi}_{\boldsymbol{R_X}}\boldsymbol{\Phi}_{\boldsymbol{R_X}}^{\boldsymbol{T}}\boldsymbol{\Phi}_{\boldsymbol{R_W}} \otimes \boldsymbol{\Phi}_{\boldsymbol{C_W}}^{\boldsymbol{T}}\boldsymbol{\Phi}_{\boldsymbol{C_X}}\boldsymbol{\Phi}_{\boldsymbol{C_X}}^{\boldsymbol{T}}\boldsymbol{\Phi}_{\boldsymbol{C_W}} \right) \\
\boldsymbol{C_{21}} &= -\boldsymbol{C_{22}}\left(\mathbf{1}_{\boldsymbol{J}} \otimes \boldsymbol{\Phi}_{\boldsymbol{R_W}}^{\boldsymbol{T}}\boldsymbol{\Phi}_{\boldsymbol{R_X}} \otimes \boldsymbol{\Phi}_{\boldsymbol{C_W}}^{\boldsymbol{T}}\boldsymbol{\Phi}_{\boldsymbol{C_X}} \right)
\end{aligned}
$$

Estimate scores using projection method

The idea is to project $\boldsymbol{Y}_{ij}$ on the space spanned by the first $K_1 \times K_2$ PCs of $\boldsymbol{X}_i$ and the first $L_1 \times L_2$ PCs of $\boldsymbol{W}_{ij}$, which reduces the dimensionality during estimation from $D \times T$ to $K_1 \times K_2$ or $L_1 \times L_2$.

Define

$$
A_{ijk_1k_2} = \phi_{C_Xk_1}^T(\mathbf{Y}_{ij} - \boldsymbol{M})\phi_{R_Xk_2} = \xi_{ik_1k_2}^X + \phi_{C_Xk_1}^T\Phi_{C_W}\boldsymbol{\Gamma}_{\boldsymbol{ij}}^{\boldsymbol{W}}\Phi_{R_W}^T\phi_{R_Xk_2} + \varepsilon_{ijk_1k_2}^{(1)}
$$

and

$$
B_{ijl_1l_2} = \phi_{C_Wl_1}^T(\mathbf{Y}_{ij} - \boldsymbol{M})\phi_{R_Wl_2} = \xi_{ijl_1l_2}^W + \phi_{C_Wl_1}^T\Phi_{C_X}\boldsymbol{\Gamma}_{\boldsymbol{i}}^{\boldsymbol{X}}\Phi_{R_X}^T\phi_{R_Wl_2} + \varepsilon_{ijl_1l_2}^{(2)}
$$

where

$$\varepsilon^{(1)}_{ijk_1k_2} = \phi^T_{C_Xk_1}\left(\sum_{l_1=L_1+1}^{\infty}\sum_{l_2=L_2+1}^{\infty}\phi_{C_Wl_1}\xi^W_{ijl_1l_2}\phi^T_{C_Wl_2}\right)\phi_{R_Xk_2}$$
$$\varepsilon^{(2)}_{ijl_1l_2} = \phi^T_{C_Wl_1}\left(\sum_{k_1=K_1+1}^{\infty}\sum_{k_2=K_2+1}^{\infty}\phi_{C_Xk_1}\xi^X_{ijk_1k_2}\phi^T_{C_Xk_2}\right)\phi_{R_Wl_2}$$

It can be proved that $\text{var}\left(\varepsilon^{(1)}_{ijk_1k_2}\right) \leq \lambda_{C_WL_1+1}\lambda_{R_WL_2+1}$ and $\text{cov}\left(\varepsilon^{(1)}_{ijk_1k_2}, \varepsilon^{(1)}_{ijk'_1k'_2}\right) \leq \lambda_{C_WL_1+1}\lambda_{R_WL_2+1}$. So if L_1 and L_2 are relatively large then we can treat $\varepsilon^{(1)}_{ijk_1k_2}$ to have a diagonal covariance matrix. A similar approximation can be applied to $\varepsilon^{(2)}_{ijl_1l_2}$. Therefore the full model can be approximated as

$$\begin{cases} A_{ijk_1k_2} = \xi^X_{ik_1k_2} + \phi^T_{C_Xk_1}\Phi_{C_W}\boldsymbol{\Gamma}^{\boldsymbol{W}}_{\boldsymbol{ij}}\Phi^T_{R_W}\phi_{R_Xk_2} + \varepsilon^{(1)}_{ijk_1k_2} \\ B_{ijl_1l_2} = \xi^W_{ijl_1l_2} + \phi^T_{C_Wl_1}\Phi_{C_X}\boldsymbol{\Gamma}^{\boldsymbol{X}}_{\boldsymbol{i}}\Phi^T_{R_X}\phi_{R_Wl_2} + \varepsilon^{(2)}_{ijl_1l_2} \\ \boldsymbol{\Gamma}^{\boldsymbol{X}}_{\boldsymbol{i}} \sim \text{MN}_{D,T}\left(0, D_{C_X}, D_{R_X}\right) \\ \boldsymbol{\Gamma}^{\boldsymbol{W}}_{\boldsymbol{ij}} \sim \text{MN}_{D,T}\left(0, D_{C_W}, D_{R_W}\right) \\ \varepsilon^{(1)}_{ijk_1k_2} \sim N(0, \tau_1^{-1}) \\ \varepsilon^{(2)}_{ijl_1l_2} \sim N(0, \tau_2^{-1}) \end{cases} \tag{A.1.1}$$

We then can estimate the scores using Monte Carlo Markov Chain (MCMC).

From Model A.1.1, we can calculate the conditional posterior distribution for within- and

between-subject PC scores as

$$\xi^X_{ik_1k_2}|A_{ijk_1k_2}, \xi^W_{ijl_1l_2} \sim \text{Normal}(\mu_{\xi^X_{ik_1k_2}}, V_{\xi^X_{ik_1k_2}})$$

$$\xi^W_{ijl_1l_2}|B_{ijl_1l_2}, \xi^X_{ik_1k_2} \sim \text{Normal}(\mu_{\xi^W_{ijl_1l_2}}, V_{\xi^W_{ijl_1l_2}})$$

where

$$\mu_{\xi^X_{ik_1k_2}} = \frac{\lambda_{C_Xk_1}\lambda_{R_Xk_2}\sum_{j=1}^{J_i}\left(A_{ijk_1k_2} - \phi^T_{C_Xk_1}\Phi_{C_W}\boldsymbol{\Gamma^W_{ij}}\Phi^T_{R_W}\phi_{R_Xk_2}\right)}{\tau_1^{-1} + J_i\lambda_{C_Xk_1}\lambda_{R_Xk_2}}$$

$$V_{\xi^X_{ik_1k_2}} = \frac{\tau_1^{-1}\lambda_{C_Xk_1}\lambda_{R_Xk_2}}{\tau_1^{-1} + J_i\lambda_{C_Xk_1}\lambda_{R_Xk_2}}$$

and

$$\mu_{\xi^W_{ijl_1l_2}} = \frac{\lambda_{C_Wl_1}\lambda_{R_Wk_2}\left(B_{ijk_1k_2} - \phi^T_{C_Wl_1}\Phi_{C_X}\boldsymbol{\Gamma^X_i}\Phi^T_{R_X}\phi_{R_Wl_2}\right)}{\tau_2^{-1} + \lambda_{C_Wl_1}\lambda_{R_Wl_2}}$$

$$V_{\xi^X_{ijl_1l_2}} = \frac{\tau_2^{-1}\lambda_{C_Wl_1}\lambda_{R_Wl_2}}{\tau_2^{-1} + \lambda_{C_Wl_1}\lambda_{R_Wl_2}}$$

Using above distributions, we then generate joint posterior distribution for both within- and between-subject PC scores by Gibbs sampler.

Detailed results for the simulation

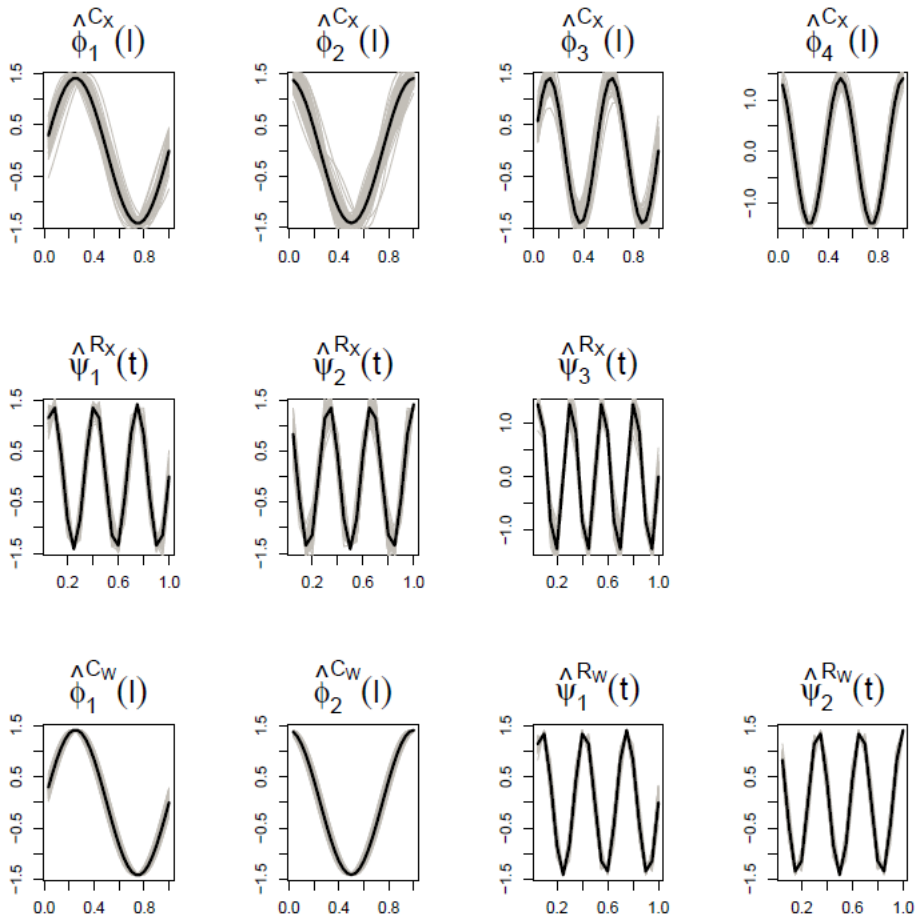

Figure A.1.1: Simulation result for signal-to-noise ratio $+\infty$.

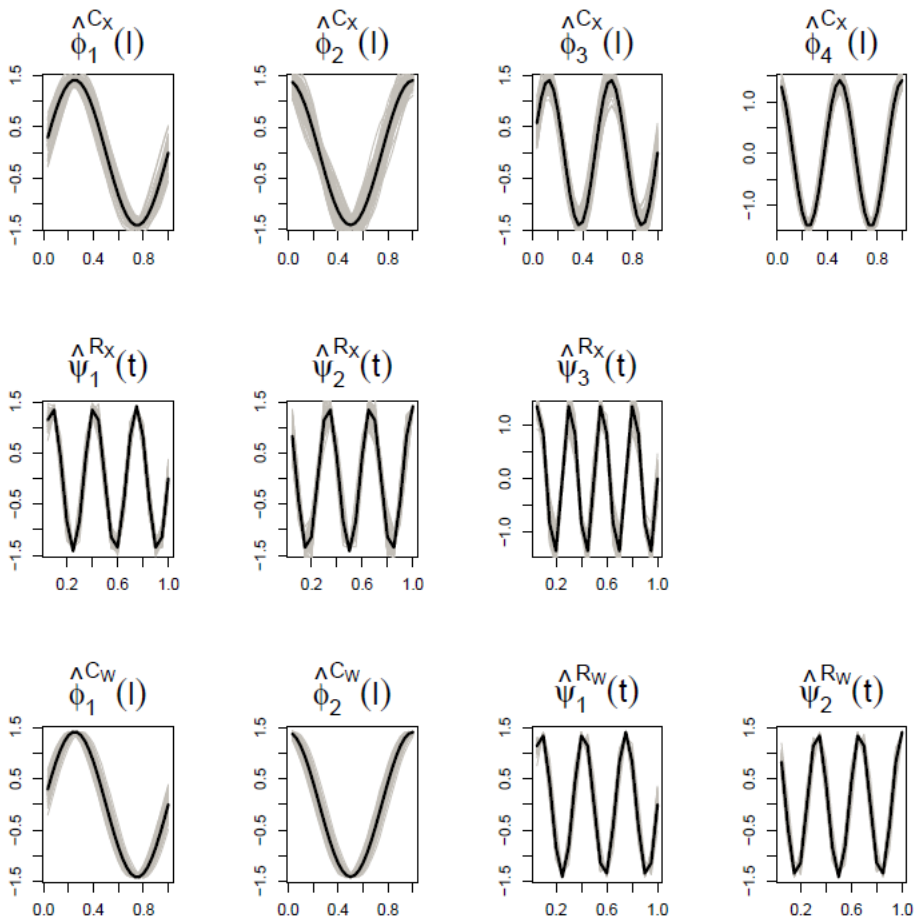

Figure A.1.2: Simulation result for signal-to-noise ratio 10.

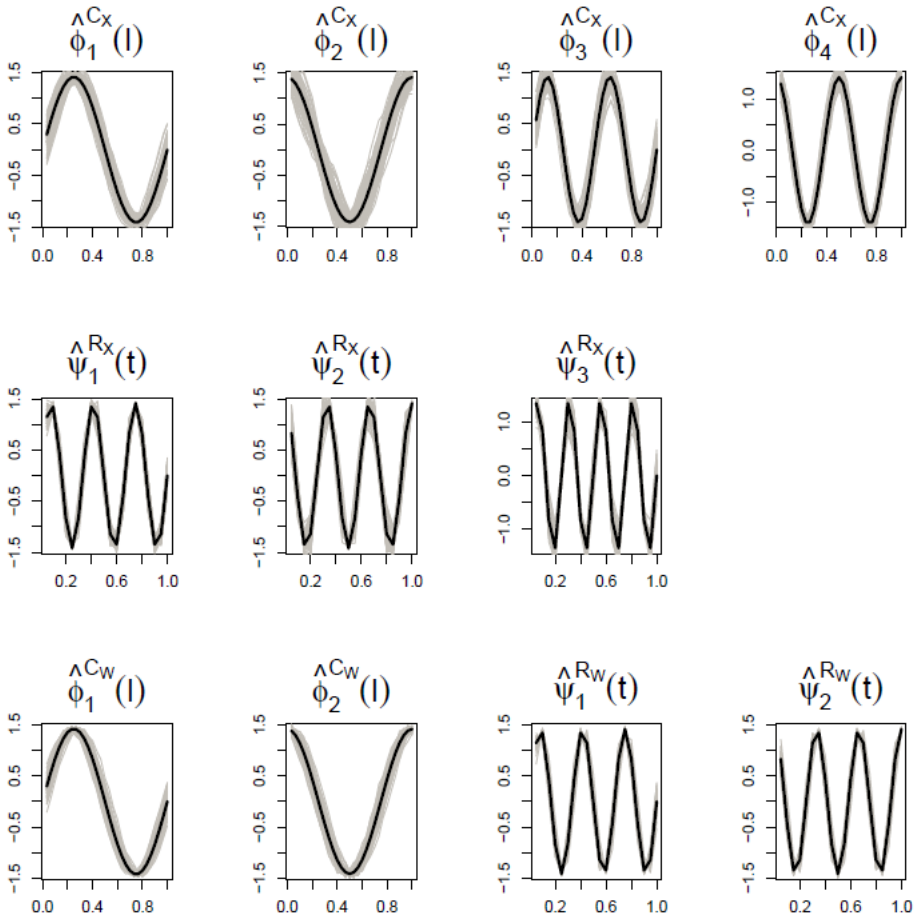

Figure A.1.3: Simulation result for signal-to-noise ratio 1.

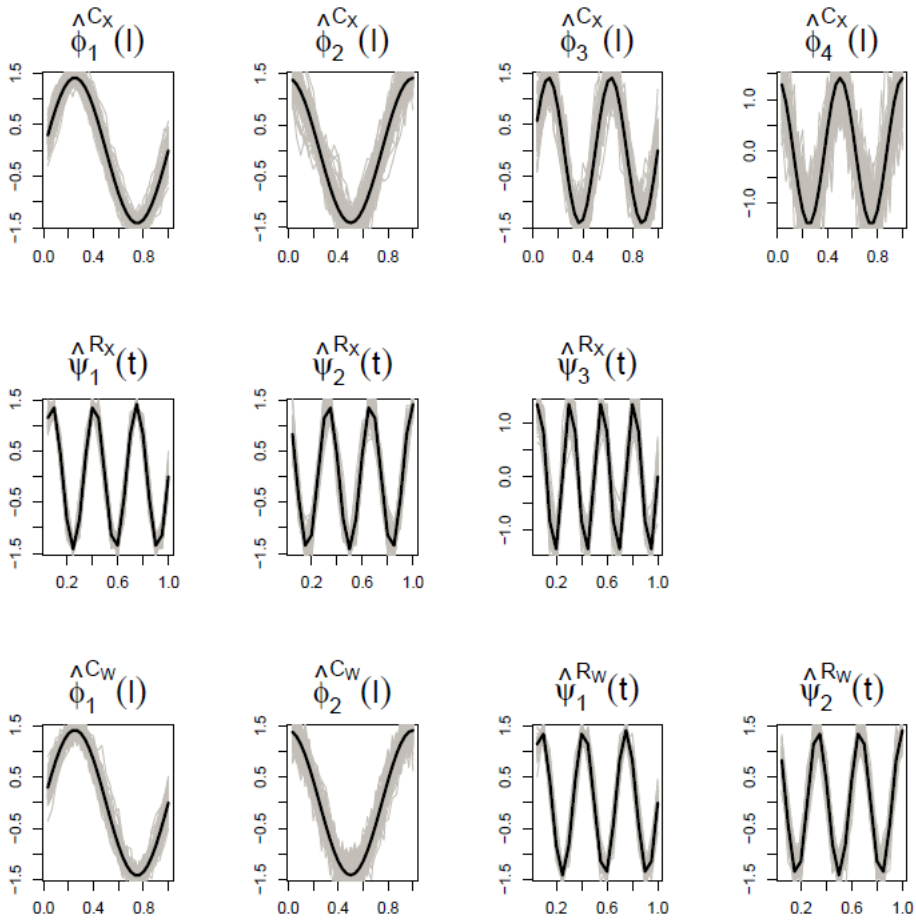

Figure A.1.4: Simulation result for signal-to-noise ratio 0.1.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Allen, G. I. (2013a). Multi-way functional principal components analysis. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2013 IEEE 5th International Workshop on*, pages 220–223. IEEE.

Allen, G. I. (2013b). Sparse and functional principal components analysis. *arXiv preprint arXiv:1309.2895*.

Allen, G. I., Grosenick, L., and Taylor, J. (2014). A generalized least squares matrix decomposition. *Journal of the American Statistical Association*, 109:145–159.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. Wiley-Interscience.

Ashburner, J. and Friston, K. (2000). Voxel-based morphometry—the methods. *NeuroImage*, 11(6):805–821.

BIBLIOGRAPHY

Atlas, L. Y., Lindquist, M. A., Bolger, N., and Wager, T. D. (2014). Brain mediators of the effects of noxious heat on pain. *Pain*, 155:1632–1648.

Barkhof, F. (2002). The clinico-radiological paradox in multiple sclerosis revisited. *Current Opinion in Neurology*, 15(3):239.

Barnea-Goraly, N., Kwon, H., Menon, V., Eliez, S., Lotspeich, L., and Reiss, A. (2004). White matter structure in autism: preliminary evidence from diffusion tensor imaging. *Biological Psychiatry*, 55(3):323–326.

Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*. Springer.

Bunea, F., She, Y., Ombao, H., Gongvatana, A., Devlin, K., and Cohen, R. (2011). Penalized least squares regression methods and applications to neuroimaging. *NeuroImage*, 55(4):1519–1527.

Caffo, B. S., Crainiceanu, C. M., Verduzco, G., Joel, S., Mostofsky, S. H., Bassett, S. S., and Pekar, J. J. (2010). Two-stage decompositions for the analysis of functional connectivity for fMRI with application to Alzheimer's disease risk. *Neuroimage*, 51(3):1140–1149.

Calhoun, V., Adali, T., Pearlson, G., and Pekar, J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Human brain mapping*, 14(3):140–151.

Carroll, M., Cecchi, G., Rish, I., Garg, R., and Rao, A. (2009). Prediction and interpretation of distributed neural activity with sparse models. *NeuroImage*, 44(1):112–122.

BIBLIOGRAPHY

Cercignani, M., Inglese, M., Pagani, E., Comi, G., and Filippi, M. (2001). Mean diffusivity and fractional anisotropy histograms of patients with multiple sclerosis. *American Journal of Neuroradiology*, 22(5):952–958.

Ciccarelli, O., Catani, M., Johansen-Berg, H., Clark, C., and Thompson, A. (2008). Diffusion-based tractography in neurological disorders: concepts, applications, and future developments. *The Lancet Neurology*, 7(8):715–727.

Cipra, B. (1987). An introduction to the Ising model. *American Mathematical Monthly*, 94(10):937–959.

Cohen, J., Asarnow, R., Sabb, F., Bilder, R., Bookheimer, S., Knowlton, B., and Poldrack, R. (2011). Decoding continuous variables from neuroimaging data: basic and clinical applications. *Frontiers in Neuroscience*, 5.

Crainiceanu, C. M., Caffo, B. S., Luo, S., and Zipunnikov, V. (2011). Population value decomposition, a framework for the analysis of image populations. *Journal of the American Statistical Association*, 106:775–790.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley-Interscience.

Dawid, P. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274.

de Brecht, M. and Yamagishi, N. (2012). Combining sparseness and smoothness improves classification accuracy and interpretability. *NeuroImage*, 60(2):1550–1561.

BIBLIOGRAPHY

Di, C., Crainiceanu, C. M., and Jank, W. S. (2014a). Multilevel sparse functional principal component analysis. *STAT*, 3(1):126–143.

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *Annals of Applied Statistics*, 3(1):458–488.

Di, C.-Z., Jank, W. S., and Crainiceanu, C. M. (2014b). Multilevel sparse functional principal component analysis. *STAT*, 3(1):126–143.

Dien, J., Spencer, K. M., and Donchin, E. (2003). Localization of the event-related potential novelty response as defined by principal components analysis. *Cognitive Brain Research*, 17(3):637–650.

Everitt, B. and Bullmore, E. (1999). Mixture model mapping of brain activation in functional magnetic resonance images. *Human Brain Mapping*, 7(1):1–14.

Fischer, J., Rudick, R., Cutter, G., Reingold, S., and Reingold (1999). The Multiple Sclerosis Functional Composite measure (MSFC): an integrated approach to MS clinical outcome assessment. *Multiple Sclerosis*, 5(4):244–250.

Flandin, G. and Penny, W. (2007). Bayesian fMRI data analysis with sparse spatial basis function priors. *NeuroImage*, 34(3):1108–1125.

Fosdick, B. K. and Hoff, P. D. (2014). Separable factor analysis with applications to mortality data. *Annals of Applied Statistics*, 8(1):120–147.

BIBLIOGRAPHY

Friston, K., Holmes, A., Worsley, K., Poline, J., Frith, C., and Frackowiak, R. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, 2(4):189–210.

Gauch, H. G. (1988). Model selection and validation for yield trials with interaction. *Biometrics*, 44:705–715.

Genton, M. G. (2007). Separable approximations of space-time covariance matrices. *Environmetrics*, 18(7):681–695.

Goldsmith, J., Crainiceanu, C., Caffo, B., and Reich, D. (2011). Penalized functional regression analysis of white-matter tract profiles in multiple sclerosis. *NeuroImage*, 57(2):431–439.

Goldsmith, J., Huang, L., and Crainiceanu, C. M. (2013). Smooth scalar-on-image regression. *Journal of Computational and Graphical Statistics*. To appear.

Goodin, D. (2006). Magnetic resonance imaging as a surrogate outcome measure of disability in multiple sclerosis: have we been overly harsh in our assessment? *Annals of Neurology*, 59(4):597–605.

Greven, S., Crainiceanu, C., Caffo, B., and Reich, D. (2010). Longitudinal functional principal component analysis. *Electronic Journal of Statistics*, 4:1022.

Harrison, D., Caffo, B., Shiee, N., Farrell, J., Bazin, P., Farrell, S., Ratchford, J., Calabresi,

P., and Reich, D. (2011). Longitudinal changes in diffusion tensor–based quantitative MRI in multiple sclerosis. *Neurology*, 76(2):179–186.

Hartvig, N. and Jensen, J. (2000). Spatial mixture modeling of fMRI data. *Human Brain Mapping*, 11(4):233–248.

Hasan, K., Gupta, R., Santos, R., Wolinsky, J., and Narayana, P. (2005). Diffusion tensor fractional anisotropy of the normal-appearing seven segments of the corpus callosum in healthy adults and relapsing-remitting multiple sclerosis patients. *Journal of Magnetic Resonance Imaging*, 21(6):735–743.

Haxby, J., Gobbini, M., Furey, M., Ishai, A., Schouten, J., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430.

Haynes, J. and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hoff, P. D. (2014). Multilinear tensor regression for longitudinal relational data. *arXiv preprint arXiv:1412.0048*.

Huang, J. Z., Shen, H., and Buja, A. (2008). Functional principal components analysis via penalized rank one approximation. *Electronic Journal of Statistics*, 2:678–695.

BIBLIOGRAPHY

Huang, J. Z., Shen, H., and Buja, A. (2009). The analysis of two-way functional data using two-way regularized singular value decompositions. *Journal of the American Statistical Association*, 104:1054.

Huettel, S. A., Song, A. W., and McCarthy, G. (2004). *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, MA.

Hung, H., Wu, P., Iping, T., and Suyun, H. (2012). On multilinear principal component analysis of order-two tensors. *Biometrika*, 99(3):569–583.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430.

Jolliffe, I. (2002). *Principal component analysis*. Springer-Verlag New York.

Keller, T., Kana, R., and Just, M. (2007). A developmental study of the structural integrity of white matter in autism. *Neuroreport*, 18:23–27.

Kern, K., Sarcona, J., Montag, M., Giesser, B., and Sicotte, N. (2010). Corpus callosal diffusivity predicts motor impairment in relapsing-remitting multiple sclerosis: A TBSS and tractography study. *NeuroImage*, 55(3):1699–1677.

Koch, G. G. (1967). A general approach to the estimation of variance components. *Technometrics*, 9(1):93–118.

Landman, B., Farrell, J., Jones, C., Smith, S., Prince, J., and Mori, S. (2007). Effects of diffusion weighting schemes on the reproducibility of DTI-derived fractional

anisotropy, mean diffusivity, and principal eigenvector measurements at 1.5 T. *NeuroImage*, 36(4):1123–1138.

Lee, M., Shen, H., Huang, J. Z., and Marron, J. (2010). Biclustering via sparse singular value decomposition. *Biometrics*, 66(4):1087–1095.

Lin, X., Tench, C., Morgan, P., and Constantinescu, C. (2008). Use of combined conventional and quantitative MRI to quantify pathology related to cognitive impairment in multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(4):437–441.

Lin, X., Tench, C., Morgan, P., Niepel, G., and Constantinescu, C. (2005). 'Importance sampling' in MS: Use of diffusion tensor tractography to quantify pathology related to specific impairment. *Journal of the Neurological Sciences*, 237(1):13–19.

Lindquist, M. A. (2008). The statistical analysis of fMRI data. *Statistical Science*, 23(4):439–464.

Lindquist, M. A. (2012). Functional causal mediation analysis with an application to brain connectivity. *Journal of the American Statistical Association*, 107:1297–1309.

Lowe, M., Horenstein, C., Hirsch, J., Marrie, R., Stone, L., Bhattacharyya, P., Gass, A., and Phillips, M. (2006). Functional pathway-defined MRI diffusion measures reveal increased transverse diffusivity of water in multiple sclerosis. *NeuroImage*, 32(3):1127–1133.

BIBLIOGRAPHY

Lu, N. and Zimmerman, D. L. (2005). The likelihood ratio test for a separable covariance matrix. *Statistics & probability letters*, 73(4):449–457.

Mitchell, M. W., Genton, M. G., and Gumpertz, M. L. (2006). A likelihood ratio test for separability of covariances. *Journal of Multivariate Analysis*, 97(5):1025–1043.

Norman, K., Polyn, S., Detre, G., and Haxby, J. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9):424–430.

Ozturk, A., Smith, S., Gordon-Lipkin, E., Harrison, D., Shiee, N., Pham, D., Caffo, B., Calabresi, P., and Reich, D. (2010). MRI of the corpus callosum in multiple sclerosis: association with disability. *Multiple Sclerosis*, 16(2):166–177.

Penny, W., Trujillo-Barreto, N., and Friston, K. (2005). Bayesian fMRI time series analysis with spatial priors. *NeuroImage*, 24(2):350–362.

Reiss, P., Huo, L., Ogden, R., Zhao, Y., and Kelly, C. (2015). Wavelet-domain regression and predictive inference in psychiatric neuroimaging. *Annals of Applied Statistics*.

Reiss, P., Mennes, M., Petkova, E., Huang, L., Hoptman, M., Biswal, B., Colcombe, S., Zuo, X., and Milham, M. (2011). Extracting information from functional connectivity maps via function-on-scalar regression. *NeuroImage*, 56(1):140–148.

Reiss, P. and Ogden, R. (2010). Functional generalized linear models with images as predictors. *Biometrics*, 66(1):61–69.

BIBLIOGRAPHY

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B*, pages 233–243.

Ryali, S., Supekar, K., Abrams, D., and Menon, V. (2010). Sparse logistic regression for whole brain classification of fMRI data. *NeuroImage*, 51(2):752.

Shabalin, A. A. and Nobel, A. B. (2013). Reconstruction of a low-rank matrix in the presence of gaussian noise. *Journal of Multivariate Analysis*, 118:67–76.

Shou, H., Zipunnikov, V., Crainiceanu, C. M., and Greven, S. (2014). Structured functional principal component analysis. *Biometrics*.

Smith, A., Cullis, B., and Thompson, R. (2001). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics*, 57(4):1138–1147.

Smith, S., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T., Mackay, C., Watkins, K., Ciccarelli, O., Cader, M., Matthews, P., and Behrens, T. (2006). Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *NeuroImage*, 31(4):1487–1505.

Spencer, K. M., Dien, J., and Donchin, E. (2001). Spatiotemporal analysis of the late ERP responses to deviant stimuli. *Psychophysiology*, 38(2):343–358.

BIBLIOGRAPHY

Staniswalis, J. G. and Lee, J. J. (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 93:1403–1418.

Tian, T., Huang, J. Z., and Shen, H. (2013). Two-way regularization for meg source reconstruction via multilevel coordinate descent. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(6):545–556.

Viviani, R., Gron, G., and Spitzer, M. (2005). Functional principal component analysis of fMRI data. *Human Brain Mapping*, 24(2):109–129.

Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W., and Kross, E. (2013). An fMRI-based neurologic signature of physical pain. *New England Journal of Medicine*, 368(15):1388–1397.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.

Woolrich, M. and Behrens, T. (2006). Variational Bayes inference of spatial mixture models for segmentation. *IEEE Transactions on Medical Imaging*, 25(10):1380–1391.

Woolrich, M., Jenkinson, M., Brady, J., and Smith, S. (2004). Fully Bayesian spatio-temporal modeling of fMRI data. *IEEE Transactions on Medical Imaging*, 23(2):213–231.

BIBLIOGRAPHY

Xiao, L., Li, Y., and Ruppert, D. (2013). Fast bivariate P-splines: the sandwich smoother. *Journal of the Royal Statistical Society: Series B*, 75(3):577–599.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590.

Ye, J. (2005). Generalized low rank approximations of matrices. *Machine Learning*, 61(1-3):167–191.

Zhang, D. and Zhou, Z.-H. (2005). (2d) 2PCA: Two-directional two-dimensional PCA for efficient face representation and recognition. *Neurocomputing*, 69(1):224–231.

Zhang, L., Shen, H., Huang, J. Z., et al. (2013). Robust regularized singular value decomposition with application to mortality data. *The Annals of Applied Statistics*, 7(3):1540–1561.

Zhou, H. and Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society: Series B*, 76:463–483.

Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552.

Zipunnikov, V., Caffo, B., Yousem, D. M., Davatzikos, C., Schwartz, B. S., and Crainiceanu, C. (2011). Multilevel functional principal component analysis for high-dimensional data. *Journal of Computational and Graphical Statistics*, 20:852–873.

BIBLIOGRAPHY

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

CURRICULUM VITAE

LEI HUANG

lhuang31@jhu.edu

615 N. Wolfe St. E3040

Baltimore, MD 21205

http://www.biostat.jhsph.edu/∼lehuang

Date of Birth: Dec 8th, 1986

Place of Birth: Shanghai, China

EDUCATION

2011 - 2016 **Johns Hopkins Bloomberg School of Public Health**, Baltimore, MD

Ph.D. in Biostatistics

Thesis title: *Statistical Methods To Analyze High-Dimensional Structured Data*

Advisor: Prof. Ciprian Crainiceanu

2008 - 2009 **Columbia University**, New York, NY

M.A. in Statistics

2004 - 2008 **Shanghai Jiao Tong University**, Shanghai, China

B.S. in Mathematics

CURRICULUM VITAE

PROFESSIONAL EXPERIENCE

06/2015 - 08/2015	**Quantitative Summer Associate**
	Morgan Stanley, New York City, NY
06/2014 - 08/2014	**Predictive Analytics**
	Pivotal Software Inc., Palo Alto, CA
2010 - 2011	**Senior Data Analyst**
	NYU Medical Center, New York University, NY

HONORS AND AWARDS

JOHNS HOPKINS UNIVERSITY

2014	Jane and Steve Dykacz Award
2014	ENAR student paper Award
2014	Joseph Zeger Travel Reimbursement Award
2013 - 2016	Sommer Scholarship
2012	PhD Examination Award

SHANGHAI JIAO TONG UNIVERSITY

CURRICULUM VITAE

2006 Li & Fung Scholoarship

2005 National Scholarship of Shanghai Jiao Tong University

2004 National Scholarship of Shanghai Jiao Tong University

PUBLICATIONS

PUBLISHED/SUBMITTED

Reiss, P. T., **Huang, L.**, and Mennes, M. (2010). "Fast function-on-scalar regression with penalized basis expansions." *International Journal of Biostatistics*, 6(1), article-28.

Huang, L., Scheipl, F., Goldsmith, J., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C., Reiss, P. (2011). "refund: Regression with functional data." *R package* version 0.1-12.

Reiss, P. T., **Huang, L.**, Cavanaugh, J. E. and Roy, A. K. (2012). "Resampling-Based Information Criteria for Best-Subset Regression." *Annals of the Institute of Statistical Mathematics*, 64(6): 1161-1186.

Reiss, P. T., Mennes, M., Petkova, E., **Huang, L.**, Hoptman, M. J., Biswal, B. B., ... and Milham, M. P. (2011). "Extracting information from functional connectivity maps via function-on-scalar regression." *NeuroImage*, 56(1): 140-148.

CURRICULUM VITAE

Reiss, P. T., and **Huang, L.** (2012). "Smoothness Selection for Penalized Quantile Regression Splines." *International Journal of Biostatistics.* 8(1).

Reiss, P. T., **Huang, L.**, Chen, Y. H., Huo, L., Tarpey, T., and Mennes, M. (2012). "Massively parallel nonparametric regression, with an application to developmental brain mapping." *Journal of Computational and Graphical Statistics*, 23(1): 232-248.

Goldsmith, J., **Huang, L.**, and Crainiceanu, C. M. (2012). "Smooth Scalar-on-Image Regression via Spatial Bayesian Variable Selection." *Journal of Computational and Graphical Statistics*, 23(1):46-64.

Reiss, P. T., Schwartzman, A., Lu, F., **Huang, L.**, and Proal, E. (2012). "Paradoxical results of adaptive false discovery rate procedures in neuroimaging studies.", *NeuroImage*, 63(4): 1833-1840.

Abikoff, H., Gallagher, R., Wells, K. C., Murray, D. W., **Huang, L.**, Lu, F., and Petkova, E. (2012). "Remediating Organizational Functioning in Children With ADHD: Immediate and Long-Term Effects From a Randomized Controlled Trial." *Journal of consulting and clinical psychology*, 81(1): 113.

Reiss, P. T., **Huang, L.** (2012). reams: Resampling-Based Adaptive Model Selection.

Petkova, E., Tarpey, T., **Huang, L.**, and Deng, L. (2013). "Interpreting meta-regression: application to recent controversies in antidepressants' efficacy". *Statistics in medicine*.

Reiss, P. T., Chen, Y. H., **Huang, L.**, and Huo, L. (2013). vows: Voxelwise Semiparametrics.

CURRICULUM VITAE

Huang, L., Goldsmith, J., Reiss, P. T., Reich, D. and Crainiceanu, C. M. (2013).“Bayesian Scalar-on-Image Regression with Application to Association Between Intracranial DTI and Cognitive Outcomes”. *NeuroImage*.

Xiao, L., **Huang, L.**, Schrack, J. A., Ferrucci, L., Zipunnikov, V. and Crainiceanu, C. M. (2015). “Quantifying the life-time circadian rhythm of physical activity: a covariate-dependent functional approach”. *Biostatistics*.

Reiss, P. T., **Huang, L.**, Chen, H., Colcombe, S. “Varying-Smoother Models for Functional Responses”, 2013. Submitted to *Journal of the Royal Statistical Society*.

Cooper, R., **Huang, L.**, Hardy, R., Harris, T., Schrack, J., Crainiceanu, C. and Kuh, D. “Associations of contemporaneous BMI and obesity history with daily patterns of physical activity at 60-64 years: findings from a British birth cohort study”, 2016. Submitted to *The American Journal of Clinical Nutrition*.

Huang, L., Reiss, P. T., Xiao, L., Zipunnikov, V., Lindquist, M. and Crainiceanu, C. M. (2016). “Two-way principal component analysis for matrix-variate data, with an application to functional magnetic resonance imaging data”, 2016. Submitted to *Biostatistics*.

WORKING PAPERS

Huang, L., Zipunnikov, V. and Crainiceanu, C. M. (2014). “Multilevel principal component analysis on matrix data”.

Huang, L., Gellar, J., Xiao, L. and Crainiceanu, C. (2015). “Dynamic prediction using historical survival model”.

CURRICULUM VITAE

TEACHING

2015 Introduction to Machine Learning, Graduate, Vadim Zipunnikov.

2014 Statistical Methods textbfI-II, Graduate, 140.651-652.

2014 Bayesian Data Analysis, Graduate, Gary Rosner.

2013 Advanced Methods in Biostatistics, Graduate, 140.755-756, Vadim Zipunnikov.

2013 Statistical Methods in Public Health **III-IV**, Graduate, 140.623-624.

2012 Statistical Methods in Public Health **I-II**, Graduate, 140.621-622.